How to Learn when Data Gradually Reacts to Your Model

Zachary Izzo
Dept. of Mathematics
Stanford University
zizzo@stanford.edu

James Zou
Dept. of Biomedical Data Science
Stanford University
jamesz@stanford.edu

Lexing Ying
Dept. of Mathematics & ICME
Stanford University
lexing@stanford.edu

Abstract

A recent line of work has focused on training machine learning (ML) models in the performative setting, i.e. when the data distribution reacts to the deployed model. The goal in this setting is to learn a model which both induces a favorable data distribution and performs well on the induced distribution, thereby minimizing the test loss. Previous work on finding an optimal model assumes that the data distribution immediately adapts to the deployed model. In practice, however, this may not be the case, as the population may take time to adapt to the model. In many applications, the data distribution depends on both the currently deployed ML model and on the "state" that the population was in before the model was deployed. In this work, we propose a new algorithm, Stateful Performative Gradient Descent (Stateful PerfGD), for minimizing the performative loss even in the presence of these effects. We provide theoretical guarantees for the convergence of Stateful PerfGD. Our experiments confirm that Stateful PerfGD substantially outperforms previous state-of-the-art methods.

1 INTRODUCTION

A recent line of work has sought to study how to effectively train machine learning (ML) models in the presence of performative effects (Perdomo et al., 2020). Performativity describes the scenario in which our deployed model or algorithm effects the distribution of the data or population which we are studying. Such effects can be expected when our model is used to make consequential decisions concerning the population. As ML becomes ever more ubiquitous across fields, considering these performative effects also grows in importance.

For example, suppose a bank uses a ML model which considers user features—e.g. income, number of open credit lines, etc.—to decide which user should be granted a loan. Based on the original data distribution, the model learns that people with more credit lines open are more likely to repay their loans. After the model is deployed, some users may open more credit lines in order to improve their chances of receiving a loan. In this case, the data distribution has changed as a direct consequence of deploying a specific model. More importantly, the distribution of the outcome—whether or not the person repays his or her loan—given the features has changed, leading to degraded model performance.

Formally, we assume that deploying a model induces a new distribution over test data. The goal of model training under performative distribution shift is to minimize *performative risk*, i.e., the model's loss on the distribution it induces. Recently, Izzo et al. (2021) proposed a "meta-algorithm" (performative gradient descent or PerfGD) to accomplish this when the induced data distribution depends only on the deployed model. This amounts to assuming that the data distribution immediately adapts to the deployed model, irrespective of any other conditions. In practice, such a model of performative effects may be overly simplistic. It is likely that the induced distribution will depend not only on the deployed model, but also some notion of the "state" that the population was in when the model was deployed. In the loan example, for instance, it will take loan applicants some time to open new credit lines, so we can expect the distribution to change gradually as applicants have more time to adapt, before finally settling to some steady-state distribution for the deployed model. Optimizing the test loss in the state-dependent performative case has been understudied in the literature, and the addition of a state (which cannot be controlled explicitly, only implicitly) increases

Proceedings of the 25$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

the difficulty of the optimization. We propose a novel algorithm and analysis to fill this important gap.

1.1 Our Contributions

In this work, we introduce a new algorithm for minimizing the performative risk in the state-dependent performative setting. Our approach is similar in spirit to that of Izzo et al. (2021), in that it amounts to estimating an appropriate gradient and using it to perform gradient descent. However, unlike Izzo et al. (2021), we no longer have even direct sample access to the distribution that we care about (the "long-term" induced distribution, i.e. the distribution which the population will finally settle to over time), and this added technical challenge makes previous algorithms for optimizing the performative risk ineffective. Indeed, the only way to apply previous approaches directly is to wait for many time steps after each model deployment so that the induced distribution stabilizes to its long-term limit. Our algorithm overcomes this limitation by "simulating" waiting, without actually needing to do so. We show theoretically that this method accurately captures the behavior of the long-term distribution. Experiments confirm our theory and also show its improvement over existing methods which are not specifically adapted to the stateful setting.

1.2 Related Work

It has long been known that changes between training and test distributions can lead to catastrophic failure for many ML models. The general problem of non-identical training and test distributions is known as distribution shift or dataset shift, and there is an extensive literature which seeks to address these issues when training models (Quionero-Candela et al., 2009; Storkey, 2009; Moreno-Torres et al., 2012). Much of the work in this area has been devoted to dealing with shifts due to *external* factors outside of the modeler's control, and developing methods to cope with these changes is still a highly active area of research (Koh et al., 2021).

Perdomo et al. (2020) proposed studying distribution shifts which arise due to the deployed model itself, referred to as performative distribution shift. They gave two simple algorithms—repeated risk minimization (RRM) and repeated gradient descent (RGD)—which converge to a stable point, i.e. a model which is optimal for the distribution it induces. Other early work in this area also explored stochastic algorithms for finding stable points (Mendler-Dünner et al., 2020; Drusvyatskiy and Xiao, 2020).

State-dependence in the performative setting was introduced by Brown et al. (2020). A notion of optimality in this setting is the minimization of the *long-term* performative loss—that is, finding a model which minimizes the average risk over an infinite time horizon, assuming that we keep deploying that same model. Brown et al. (2020) showed that the RRM procedure introduced by Perdomo et al. (2020) converges to long-term stable points. RRM and RGD rely on population-level quantities (e.g., minimization of the population-level risk or a population-level gradient). Li and Wai (2021) extended these results to show that stochastic optimization algorithms also find a performatively stable point in the stateful setting. We remark that these works differ from ours in that they both seek to find a stable point rather than an optimal point (i.e. one which minimizes the test loss), and in general stable points can be far from optimal (Izzo et al., 2021; Miller et al., 2021).

Izzo et al. (2021) proposed a method (PerfGD) for computing the performative optimum in the non-state-dependent case. Under parametric assumptions on the performative distribution, they show how to construct an approximate gradient of the performative loss and then use it to perform gradient descent. Miller et al. (2021) also studied optimizing the (stateless) performative loss. The authors quantified when the performative loss is convex and proposed using black-box derivative-free optimization methods to find the performative optimum. For certain classes of performative effects, they also propose a model-based approach to minimizing the performative loss.

A related line of work studies the setting of strategic classification (Hardt et al., 2016), which is a subclass of the general performative setting. In this setting, it is assumed that individual datum react to a deployed model by a best-response mechanism, inducing a population-level distribution shift. Dong et al. (2018) considered optimizing the performative risk in an online version of this problem and for a certain class of best-response dynamics. Other recent work in this area includes developing practically useful tools for modeling strategic behavior, such as differentiable surrogates for strategic responses and regularizers for inducing socially advantageous strategic responses (Levanon and Rosenfeld, 2021); incorporating more realistic limitations on the best-response behavior of the agents (Ghalme et al., 2021; Jagadeesan et al., 2021); examining the effects of the relative frequency of updates between the modeler and the agents (Zrnic et al., 2021); and studying the statistical and computational complexity of strategic classification in a PAC framework (Sundaram et al., 2021). While strategic classification offers a wealth of important examples of performative effects, the performative setting is more general as the change in the data need not arise from a best-response mechanism.

The original performative optimization problem can be framed as a derivative-free optimization (DFO) (Flaxman et al., 2005) problem with a noisy function value oracle. In the stateful case, however, we no longer even have an unbiased noisy oracle for the function we wish to optimize (the long-term performative risk), making black-box DFO algorithms ineffective.

2 PROBLEM SETUP

We refer readers unfamiliar with the performative literature to the introductory sections of Perdomo et al. (2020) and Izzo et al. (2021) for a complete discussion of the original (stateless) performative prediction setting.

We consider a generalization of the performative prediction problem (Perdomo et al., 2020), introduced by Brown et al. (2020) and referred to as "stateful" performativity. Let Θ denote the set of admissible model parameters and $\mathcal{Z}$ denote the data sample space. We assume that there is a *distribution map* $\mathcal{D}: \Theta \times \mathcal{M}(\mathcal{Z}) \to \mathcal{M}(\mathcal{Z})$, where $\mathcal{M}(\mathcal{Z})$ denotes the set of probability measures on $\mathcal{Z}$. If ρ_t denotes the data distribution at time t and θ_t denotes the model that we deploy at time t, then we have

$$\rho_t = \mathcal{D}(\theta_t, \rho_{t-1}).$$

That is, the data distribution at time t is a function of the model we deployed, as well as the previous state that the population was in (encoded by the previous distribution ρ_{t-1}). Note that this setting is strictly more general than the original setup of Perdomo et al. (2020), in which $\rho_t = \mathcal{D}(\theta_t)$ depended only on the deployed model. This generalization captures the fact that in practice, it is unlikely that the population we are modeling will immediately snap to a new distribution upon deployment of a new model. In general, it will take the distribution some time to adapt.

Under reasonable regularity conditions on $\mathcal{D}$, if we define $\theta_t \equiv \theta$ for all t, then there exists a limiting distribution $\rho^*(\theta) = \lim_{t\to\infty} \rho_t$. (See Claim 1 of Brown et al. (2020) for sufficient conditions.) That is, $\rho^*(\theta)$ describes the limiting distribution if we continue to deploy θ for all time steps t, and it is assumed that this distribution is independent of the initial state. If we define the long-term performative loss $\mathcal{L}^*(\theta) = \mathbb{E}_{\rho^*(\theta)}[\ell(z;\theta)]$, then a sensible goal is to compute the long-term optimum

$$\theta_{\text{OPT}} \overset{\Delta}{=} \operatorname*{argmin}_{\theta\in\Theta} \mathcal{L}^*(\theta).$$

This is similar to the problem addressed in Izzo et al. (2021), except now we do not even have direct sample access to $\rho^*(\theta)$.

Throughout the paper, we will assume that ρ_t belongs to a parametric family with (unknown) parameter μ_t and corresponding density $p(\cdot, \mu_t)$. For concreteness, one may think of the distribution as a mixture of Gaussian with fixed covariances Σ but unknown means μ, but we remark that our techniques should be viewed more as a "meta-algorithm" whose details can be directly applied to other parametric distributions. In this setting, rather than the distribution map $\mathcal{D}$, we can equivalently consider the parameter map m, where $\mu_t = m(\theta_t, \mu_{t-1})$, and then ρ_t corresponds to the parametric distribution with parameter μ_t (i.e., ρ_t has density $p(\cdot, \mu_t)$). Analogously to the long-term distribution assumption, we will assume that for every fixed θ and any starting μ, there is a long-term parameter $\mu^*(\theta) = \lim_{k\to\infty} m^{(k)}(\theta, \mu)$, where $m^{(0)}(\theta, \mu) = \mu$ and $m^{(k)}(\theta, \mu) = m(\theta, m^{(k-1)}(\theta, \mu))$ for $k \geq 1$. That is, $m^{(k)}(\theta, \mu)$ denotes the distribution parameters after model θ has been deployed for k steps, starting from the distribution with parameters μ. For simplicity, we will assume that the model parameters θ as well as the distribution parameters μ are both d-dimensional vectors: θ, $\mu \in \mathbb{R}^d$, but we emphasize that this is for notational convenience and is not required.

Algorithm 1 describes the interaction model for our problem in terms of the parameterized distribution. Here we have assumed that, given a sample $Z = \{z_i\}_{i=1}^n$ from the distribution with parameter μ, there is some method (e.g. maximum likelihood) for estimating $\hat{\mu}(Z)$. Since there is a large literature on parametric inference, we consider $\hat{\mu}$ as provided.

Algorithm 1 Deployment and sampling model

procedure DEPLOY&SAMPLE(θ_t, μ_{t-1})
    Deploy θ_t
    Population reacts: $\mu_t \leftarrow m(\theta_t, \mu_{t-1})$
    Collect samples: $Z_t \leftarrow \{z_i^{(t)}\}_{i=1}^{n_t}$, $z_i^{(t)} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\mu_t}$
    Estimate μ_t: $\hat{\mu}_t \leftarrow \hat{\mu}(Z_t)$
    return $\hat{\mu}_t$
end procedure

Lastly, we will use $\partial_i f$ to denote the derivative of a function f with respect to its i-th argument. So for instance, $\partial_1 m(\theta, m(\theta, \mu_t))$ means the derivative of $m(\theta, m(\theta, \mu_t))$ only with respect to the θ appearing in the first argument (before the comma) even though θ appears in m in the second argument as well.

2.1 Why is the State-Dependent Case More Challenging?

As mentioned above, adding state-dependence to the performative dynamics presents a much more realistic model of performative effects likely to arise in reality. However, this added realism makes the optimization problem significantly more challenging. Indeed, we no longer even have direct access to an unbiased estimate for the function that we wish to minimize—the long-term performative loss—, as we cannot observe the long-term performative loss simply by deploying our model once. The increase in problem complexity is akin the gap between bandit problems and reinforcement learning/Markov decision processes. Thus although our setting may seem similar to that of Izzo et al. (2021) at face value, the state-dependent case is in fact a highly nontrivial advancement both in terms of the practical validity of the model and the technical/theoretical difficulty of solving the problem. Therefore, the problem demands novel algorithms and analysis, which we introduce here.

3 STATEFUL PERFGD

Our approach is to estimate the (long-term) performative gradient and then use this estimate to do approximate gradient descent. In an ideal world, we wish to compute the gradient

$$\begin{aligned}\nabla_\theta \mathcal{L}^*(\theta) &= \nabla_\theta \left[\int \ell(z;\theta)p(z;\mu^*(\theta))\,dz\right] \\ &= \int \nabla_\theta \ell(z;\theta)p(z;\mu^*(\theta))\,dz \\ &+ \int \ell(z;\theta)\frac{d\mu^*}{d\theta}^\top \nabla_\mu p(z;\mu^*(\theta))\,dz.\end{aligned}$$

(Note: $\nabla_\mu p$ denotes the gradient of the density p with respect to its μ argument. In terms of the ∂_i notation, we have $\nabla_\mu p = \partial_2 p^\top$.) There are two unknown quantities in this expression: $\mu^*(\theta)$ and $\frac{d\mu^*}{d\theta}$. The various subroutines in the algorithm are all aimed at estimating these unknown quantities. As we no longer have direct sample access to the long-term distribution $\mu^*(\theta)$, the steps needed to estimate the long-term performative gradient are different from Izzo et al. (2021), and the error analysis is more involved. There are two main components in the algorithm. First, we use the most recent H steps in the training trajectory to estimate the derivatives of the update function m (Algorithm 2), which can then be used to estimate $\frac{d\mu^*}{d\theta}$ (Equation (2)). With this estimate in hand, we can compute an estimate of the total gradient of the long-term loss and take a gradient descent step (Equation (1) and Algorithm 3). The precise steps for the algorithm are given below. In the following, $\psi_t = [\theta_t^\top,\, \mu_{t-1}^\top]^\top$ denotes the full input to m at time t, and for any collection of vectors v_i, $v_{i:j}$ denotes the matrix with columns $v_i, v_{i+1}, \ldots, v_j$.

The other estimation functions EstLTJac and EstLTGrad are given by

Algorithm 2 Estimating $\partial_i m$

Require: Estimation horizon H
- **procedure** EstPartials($\psi_{t-H:t}$, $\hat\mu_{t-H:t}$)
  - $\Delta\psi \leftarrow \psi_{t-H:t-1} - \psi_t \mathbf{1}_H^\top$
  - $\Delta\mu \leftarrow \hat\mu_{t-H:t-1} - \hat\mu_t \mathbf{1}_H^\top$
  - $[\widehat{\partial_1 m},\, \widehat{\partial_2 m}] \leftarrow (\Delta\mu)(\Delta\psi)^\dagger$
  - **return** $\widehat{\partial_1 m}$, $\widehat{\partial_2 m}$
- **end procedure**

Algorithm 3 Stateful PerfGD

Require: Estimation horizon H, perturbation size σ^2, learning rate η
- Initialize for d steps and record $\hat\mu_{t-1}$, θ_t, and $\hat\mu_t$
- **while** not converged **do**
  - $\hat\mu_t \leftarrow$ Deploy&Sample(θ_t, μ_{t-1})
  - $\widehat{\partial_1 m}, \widehat{\partial_2 m} \leftarrow$ EstPartials($\psi_{t-H:t}$, $\hat\mu_{t-H:t}$)
  - $\widehat{\frac{d\mu^*}{d\theta}} \leftarrow$ EstLTJac($\widehat{\partial_1 m}$, $\widehat{\partial_2 m}$)
  - $\widehat{\nabla\mathcal{L}_t^*} \leftarrow$ EstLTGrad(θ_t, $\hat\mu_t$, $\widehat{\frac{d\mu^*}{d\theta}}$)
  - $\theta_{t+1} \leftarrow \theta_t - \eta(\widehat{\nabla\mathcal{L}_t^*} + g_t)$, $g_t \sim \mathcal{N}(0, \sigma^2 I)$
  - $t \leftarrow t+1$
- **end while**

$$\text{EstLTJac}(\widehat{\partial_1 m}, \widehat{\partial_2 m}) = (I - \widehat{\partial_2 m})^{-1}\widehat{\partial_1 m} \tag{1}$$

$$\text{EstLTGrad}(\theta_t,\, \hat\mu_t,\, \widehat{\frac{d\mu^*}{d\theta}}) = \tag{2}$$

$$\int \nabla_\theta \ell(z;\theta_t)p(z;\hat\mu_t)\,dz + \int \ell(z;\theta_t)\widehat{\frac{d\mu^*}{d\theta}}^\top \nabla_\mu p(z;\hat\mu_t)\,dz$$

Next, we give the basic motivation for each step of this algorithm. Algorithm 2 estimates the derivatives of m using finite difference approximations gathered from the optimization trajectory so far. The columns $\Delta\psi$ are the differences in the input of m, and the columns of $\Delta\mu$ are the corresponding differences in the output of m. We then estimate the derivatives of m by solving the matrix equation $(\Delta\text{output}) \approx (\text{Jacobian})\cdot(\Delta\text{input})$. The estimation horizon H is a hyperparameter which should be tuned via standard techniques. In our proofs, we require that H be polynomially larger than the dimension d, but in practice we find that choosing $H \in [2d, 3d]$ or just using the entire previous trajectory for this step works well. We also note that $H \geq 2d$ should be enforced so that the "input difference matrix" $\Delta\psi \in \mathbb{R}^{2d\times H}$ will have a right inverse.

The formula for EstLTJac arises from the recursive definition of $m^{(k)}$ (see Section 2). Taking a derivative with respect to θ, unrolling the recursion, and sending $k \to \infty$ leads to the formula (2).

The formula for EstLTGrad is derived by taking a derivative of the long-term performative loss, recalling that this is an expectation with respect to the *known* density p with unknown parameter μ^*. As we do not know μ^* or its Jacobian $\frac{d\mu^*}{d\theta}$, we simply subsitute our best approximations for each of these ($\hat{\mu}_t$ and $\widehat{\frac{d\mu^*}{d\theta}}$, respectively) to obtain the formula (1). In Section 4, we bound the error of our approximation and show that it vanishes as the number of steps increases and as the error in $\hat{\mu}$ goes to 0. This yields an estimate for the long-term performative gradient, which we then use to take an approximate gradient descent step. The Gaussian perturbations g_t are a technical necessity which borrows ideas from smoothed analysis (Sankar et al., 2006). They ensure that the optimization trajectory has traveled enough in each direction so that the derivatives of m can be estimated even in the presence of errors in $\hat{\mu}$ and can often safely be omitted in practice. See Appendix A for a full derivation of the algorithm.

3.1 Performativity through Low-Dimensional Statistics

Performativity in ML is primarily concerned with changes in human populations as the result of a deployed model. Unless the population being model consists mostly of data scientists, it is unlikely that the constituent individuals will have a reaction based on the particular parameters of the model. Instead, individuals (and therefore the distribution of the population on the whole) likely modify their behavior based on a low-dimensional proxy, such as a credit score or classification probability. If it is the case that the distribution shift depends on a low-dimensional statistic, then we can still apply stateful PerfGD for a very high-dimensional model (e.g. a neural network) without incurring a large error due to the high dimension.

We formalize this intuition as follows. Suppose that the stateful parameter map actually takes the form $\mu_t = m(\theta_t, \mu_{t-1}) = \bar{m}(s(\theta_t, \mu_{t-1}), \mu_{t-1})$, where s is a *known* score function with $s(\theta, \mu) \in \mathbb{R}^{d_s}$ and $d_s \ll \dim(\theta)$. In this case, we may estimate the partials of $\bar{m}$ with respect to s and then use the chain rule to compute the partial of m with respect to θ, yielding $\partial_1 m(\theta_t, \mu_{t-1}) = \partial_1 \bar{m}(s_t, \mu_{t-1})\, \partial_1 s(\theta_t, \mu_{t-1})$, where $s_t = s(\theta_t, \mu_{t-1})$. Note that since s is known as a function of θ and μ, computing $\partial_1 s(\theta_t, \mu_{t-1})$ just requires estimating μ_{t-1} (which we have assumed is easy) and we instead need only estimate the derivative $\partial_1 \bar{m}$. This is a derivative with respect to d_s variables, whereas in general for this step we must compute a derivative with respect to $\dim(\theta)$ variables. When $d_s \ll \dim(\theta)$, this can make the derivative estimation task signficantly easier. We can then plug this estimator for $\partial_1 m$ into Algorithm 3 and proceed as usual.

4 THEORETICAL GUARANTEES

In this section, we quantify the performance of Stateful PerfGD theoretically. We require the following:

1. $\|\partial_1 m(\theta, \mu)\| \le B$
2. $\|\partial_2 m(\theta, \mu)\| \le \delta < 1$
3. $\sigma_{\min}(\partial_2 m(\theta, \mu)) \ge \alpha$
4. $|\ell(z; \theta)|$, $\|\nabla_\theta \ell(z; \theta)\| \le \ell_{\max}$
5. $\|\nabla^2 m\| \le C$, where $\nabla^2 m$ is the tensor of second derivatives of m, and $\|\nabla^2 \mathcal{L}^*(\theta)\| \le L$.
6. $\mathrm{diam}(\{\mu\}) \le D$, where $\{\mu\}$ denotes the set of all possibile stateful distribution parameters.
7. The estimator $\hat{\mu}$ satisfies $\|\hat{\mu}_t - \mu_t\| \le \varepsilon$.

For simplicity we will also assume that the stateful performative distribution is a Gaussian with unknown mean, i.e. the distribution parameters μ are just the mean of the Gaussian and $p(z, \mu)$ denotes the Gaussian density of z. We will also assume that the covariance is fixed and nondegenerate. The Gaussian assumption simplifies some of the already extensive calculations, but we remark that all of the results still hold for any continuous distribution with sufficiently light (e.g. sub-Gaussian) tails and smooth dependence on the distribution parameter.

With the exception of Assumptions 2 and 3, the above are all standard smoothness assumptions (Perdomo et al., 2020; Brown et al., 2020; Izzo et al., 2021; Miller et al., 2021). Assumption 2 is a sufficient condition to guarantee that $m^{(k)}(\theta, \mu) \to \mu^*(\theta)$ independent of μ, and, when combined with Assumption 6, gives us a bound on the speed of this convergence. On the other hand, Assumption 3 ensures that we are able to perturb μ_t by perturbing θ_t; without this, estimating $\partial_2 m$ will be impossible. Finally, we remark that Assumption 7 can easily be converted into a high-probability statement depending on the size of the sample collected at each step. For instance, in the case of a Gaussian mean, we have $\varepsilon = \mathcal{O}(\sqrt{\log(T)/n})$ by a simple Gaussian concentration/union bound argument.

In all of the following statements, $\mathcal{O}$ hides depedence on any of the problem-specific constants introduced in the assumptions, as well as dependence on the problem dimension and the failure probability γ. Our concern is how the error of the method behaves as the time horizon $T \to \infty$ and the estimation error $\varepsilon \to 0$. $\widetilde{\mathcal{O}}$ hides these same constants as well as log factors in T. Our main theoretical result is the following convergence theorem for Stateful PerfGD:

Theorem 1. *Let T be the number of deployments of Stateful PerfGD, and for each t let $\nabla\mathcal{L}_t^* = \nabla\mathcal{L}^*(\theta_t)$. Then for any $\gamma > 0$, there exist intervals $[\eta_{\min}, \eta_{\max}]$ and $[\sigma_{\min}, \sigma_{\max}]$ (which depend on T and the estimation error ε) such that for any learning rate η in the former and perturbation size σ in the latter interval, with probability at least $1-\gamma$, the iterates of Stateful PerfGD satisfy*

$$\min_{1\le t\le T} \|\nabla\mathcal{L}_t^*\|^2 = \widetilde{\mathcal{O}}(T^{-1/5} + \varepsilon^{1/5}).$$

Theorem 1 shows that Stateful PerfGD finds an approximate critical point. As Stateful PerfGD can be viewed as instantiating gradient descent on the long-term performative loss, and gradient descent is known to converge to minimizers (Lee et al., 2016), Stateful PerfGD will converge to an approximate local minimum. In the case that the long-term performative is convex, Stateful PerfGD will converge to the optimal point.

While the full proof of Theorem 1 requires extensive calculation, the structure of the proof is intuitive and we outline it below. We begin by bounding the error of the finite difference approximation in Algorithm 2.

Lemma 2. *Suppose that $\|\hat{\nabla}\mathcal{L}_s^*\|$ are bounded by a constant for $s < t$. Let $\frac{dm}{d\psi}$ denote the true Jacobian of m with respect to its input, and let $\widehat{\frac{dm}{d\psi}} = [\widehat{\partial_1 m}, \widehat{\partial_2 m}]$ denote the estimator from Algorithm 2. Then for appropriate choices of η, σ, and H, we have*

$$\left\|\widehat{\frac{dm}{d\psi}} - \frac{dm}{d\psi}\right\| \le \widetilde{\mathcal{O}}\left(\frac{\eta}{\sigma} + \frac{\varepsilon}{\eta\sigma^2}\right) \equiv \mathbf{E}_m.$$

Here, a smaller step size results in smaller error from the finite difference approximation, but magnifies any error in $\hat{\mu}$. Next, we analyze how the error on our estimate of the short-term derivatives translates to error on our estimates of the long-term derivatives.

Lemma 3. *The long-term Jacobian estimate $\widehat{\frac{d\mu^*}{d\theta}}$ from Eq. 2 satisfies*

$$\left\|\widehat{\frac{d\mu^*}{d\theta}} - \frac{d\mu^*}{d\theta}\right\| = \widetilde{\mathcal{O}}(\eta + \mathbf{E}_m) \equiv \mathbf{E}_{\mu^*},$$

where $\mathbf{E}_m$ is the upper bound on the error in estimating the Jacobian of m from Lemma 2.

Note that the error on the long-term Jacobian estimate also depends on the distances $\|\mu_t - \mu^*(\theta_t)\|$. A smaller learning rate gives the distribution time to adapt *during* training but without needing to wait, making these distances shrink. This can be thought of as similar to multiscale considerations in the study of PDEs (E, 2011). Next, we show that the estimation errors on $\widehat{\frac{d\mu^*}{d\theta}}$ and $\hat{\mu}_t$ remain small when we use them to estimate $\nabla\mathcal{L}_t^*$.

Lemma 4. *The estimator $\widehat{\nabla\mathcal{L}_t^*}$ from Eq. (1) satisfies*

$$\|\widehat{\nabla\mathcal{L}_t^*} - \nabla\mathcal{L}_t^*\| = \widetilde{\mathcal{O}}(\eta + \mathbf{E}_{\mu^*}),$$

where $\mathbf{E}_{\mu^*}$ is the error bound on $\widehat{\frac{d\mu^*}{d\theta}}$ from Lemma 3.

We show that the errors $\mathbf{E}_m$ and $\mathbf{E}_{\mu^*}$ from Lemmas 2 and 3 vanish at a polynomial rate as $T \to \infty$ and $\varepsilon \to 0$, so that the error in our long-term loss gradient also vanishes as the step size decreases. Finally, we use a standard analysis of gradient descent on L-smooth functions which allows for error in the gradient oracle.

Lemma 5. *Let h be any L-smooth function and let $\widehat{\nabla h}$ be a gradient oracle with bounded error: $\|\widehat{\nabla h}(x) - \nabla h(x)\| \le \mathbf{e}$, and assume that $\mathbf{e} = o(1)$. Then for η sufficiently small, the iterates x_t of gradient descent with gradient oracle $\widehat{\nabla h}$ satisfy*

$$\min_{1\le t\le T} \|\nabla h(x_t)\|^2 = \mathcal{O}\left(\frac{1}{T\eta} + \mathbf{e}\right).$$

Combining Lemmas 3-5 yields a set of dependencies on η, σ, ε, and T which can be balanced to prove Theorem 1. All of the proofs can be found in Appendix B.

5 EXPERIMENTS

In this section, we conduct experiments for all of the relevant methods, showing Stateful PerfGD's improvements over existing algorithms. First, we discuss the algorithms against which we will compare.

5.1 Previous Algorithms

Repeated Gradient Descent (RGD) This method was introduced by Perdomo et al. (2020) and refers to simply taking a gradient of the loss assuming that the distribution is fixed, then updating the model with a gradient descent step and redeploying. Li and Wai (2021) showed that RGD converges to a stable point in the long run in the stateful performative setting. Since the stateless performative problem is a subclass of the stateful one, there are cases where a stable point can be arbitrarily far from an optimal point. (See §2.2 of Izzo et al. (2021).)

PerfGD (PGD) If we repeatedly deploy each model θ until the induced distribution settles to its long-term state, then we can directly apply PerfGD from Izzo et al. (2021). While this method will eventually converge if we wait long enough at each step, we will have to deploy many suboptimal models if the induced distribution takes a long time to settle, leading to losses for the user.

Black-Box Derivative-Free Optimization (DFO) Black-box DFO seeks to optimize a function given only a function value oracle and no direct access to gradients or higher-order derivatives of the function to be optimized (Flaxman et al., 2005). The non-stateful performative prediction setting is a special case of this general problem, and black-box DFO algorithms can obtain reasonable results for non-stateful performative prediction (Miller et al., 2021). In the stateful setting, however, we no longer have a function value oracle for the long-term performative loss, so we expect black-box DFO methods to have degraded performance (if they work at all). We could take the same approach as mentioned above with PerfGD, i.e. deploying each model many times until the distribution settles to its long-term state. We note that since this method deploys perturbed versions of its best internal estimate, the cost in terms of suboptimal model deployments can be even greater than that incurred by adapting PerfGD to the stateful setting.

In all of the following figures, the solid lines denote the mean of the reported statistic and the shaded error regions denote the standard error of the mean. OPT denotes the long-term optimal loss, STAB denotes the loss of the performatively stable point, and SPGD denotes Stateful PerfGD. For details on the specific constants and hyperparameters, refer to Appendix C.

5.2 Linear m

We begin with a simple case with a linear point loss $\ell(z, \theta) = -z^\top \theta$. The long-term performative loss is $\mathcal{L}^*(\theta) = -\mu^*(\theta)^\top \theta$. We take the mean update function to be $m(\theta, \mu) = \delta\mu^*(\theta) + (1-\delta)\mu$ and set $\mu^*(\theta) = A\theta + b$ for some fixed $\delta \in (0, 1)$, a fixed matrix A and a fixed vector b. When $A \prec 0$, the long-term optimal point can be computed exactly as $\theta_{\text{OPT}} = -\frac{1}{2}A^{-1}b$.

Figure 1 compares the performance of SPGD with the other algorithms as the "amount of statefulness" varies. The x-axis is the number of deployments required before 99% of the effect of the previous mean has been removed, which corresponds to a particular δ. (If θ is deployed for k steps starting from distribution mean μ, then the mean is $(1 - (1 - \delta)^k)\mu^*(\theta) + (1 - \delta)^k\mu$, so we want $(1 - \delta)^k = 0.01$ or $\delta = 0.01^{1/k}$.) The y-axis shows the best (over a grid search of hyperparameters for each method) final performance as a fraction of OPT achieved by each of the methods after 50 model deployments. Note that since OPT is negative, a lower final loss corresponds to a larger fraction of OPT.

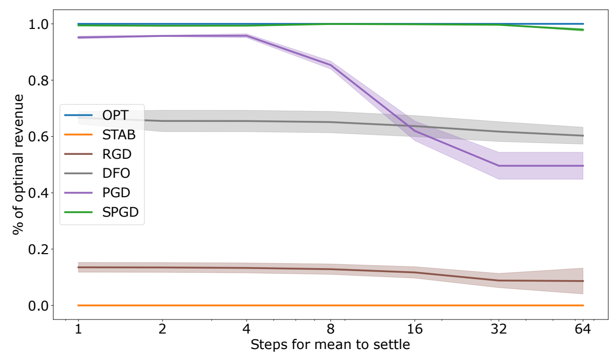

Figure 1: Fraction of optimal performance obtained by each method (higher is better, blue line is the best possible). SPGD is able to reach OPT even when the short-term mean is highly state-dependent. The other methods fail to find OPT, and their performance degrades as the statefulness of the problem increases.

While PGD and DFO make some progress towards OPT, their performance suffers even in the presence of mild state-dependence and continues to degrade as the "statefulness" of the dynamics increases. These methods must choose between longer wait times or larger errors in estimating the long-term distribution. By specifically accounting for the state dependence of the problem, SPGD maintains near-optimal performance even as the distribution takes longer to settle. For comparison, in the setting with $k = 64$ (the right-most point in Figure 1), the distribution takes 64 steps to settle, but we have only allowed 50 deployments for optimization. By *simulating* rather than *waiting* for the distribution to adapt, SPGD still reaches a near-optimal point quickly.

5.3 Nonlinear m

We alter the first example so that the rate of convergence to the long-term mean depends on the current mean and varies by coordinate. In particular, we take $m(\theta, \mu)[i] = \delta^{\mu[i]^2}\mu^*(\theta)[i] + (1 - \delta^{\mu[i]^2})\mu[i]$, with $\mu^*(\theta) = A\theta + b$ as before. Here $v[i]$ denotes the i-th component of a vector v. The long-term performative loss and optimal point are the same as before since $m^{(k)}(\theta, \mu) \to A\theta + b$, but $\partial_i m$ are more challenging to estimate.

Refer to Figure 2. Here we plot the results for a fixed δ so that we can see the training dynamics within a given scenario. The x-axis is the training iteration and the y-axis is the test loss at that iteration. In spite of the increased complexity in the derivatives of m, we see that SPGD manages to find θ_{OPT}, while the other methods have poorer performance.

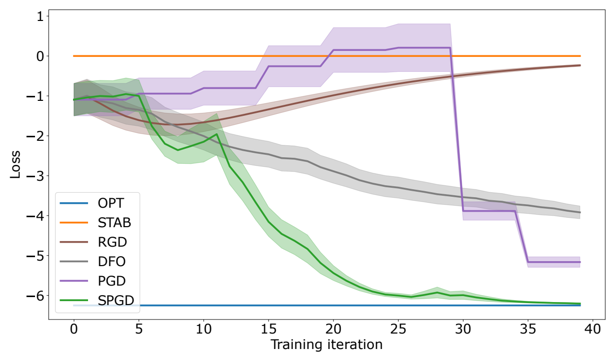

Figure 2: Results for nonlinear map m. SPGD is able to converge to θ_{OPT}, while the other methods fail to cope with the state-dependence.

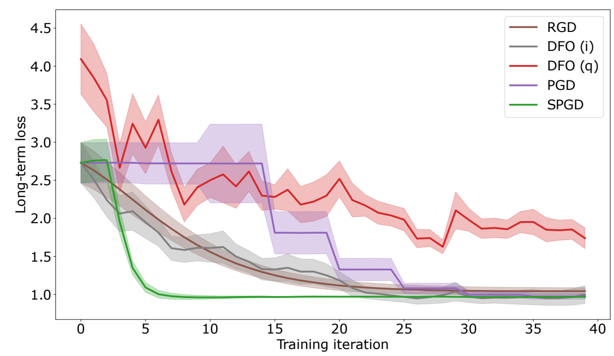

Figure 3: Results for spam classification. DFO (i) denotes the internal estimate of DFO, while DFO (q) denotes the models which are actually deployed by this algorithm. While DFO and PGD both find the optimal model, SPGD converges to it more rapidly. DFO must also deploy perturbed models (red curve) in order to find a good internal estimate. RGD converges to a stable point, resulting in $\sim 10\%$ higher final loss.

5.4 Classification

We next consider a more realistic spam classification simulation which was studied by Izzo et al. (2021). The dynamics for this experiment arise when the spammers behave strategically according to the following (state-dependent) cost function. Each spammer has some original message, denoted by the features x^{orig}, that they would like to send. This should not be thought of as an actual saved message, but rather encoding the information (e.g. a virus, scam, etc.) that they want to deliver to their victims. Their message also has a current form, denoted by the features x^{cur}. We follow the strategic classification framework (Hardt et al., 2016), where each spammer updates their message by maximizing their utility minus a modification cost, given by

$$\max_x \underbrace{-x^\top \theta}_{\text{Utility}} - \underbrace{\frac{\alpha}{2}\|x - x^{\text{orig}}\|^2}_{\text{Long-term cost}} - \underbrace{\frac{\beta}{2}\|x - x^{\text{cur}}\|^2}_{\text{Short-term cost}}.$$

The utility corresponds to the spammers' desire to receive a negative (non-spam) classification from our deployed logistic model. If we take $\alpha = \varepsilon^{-1}$ and $\beta = \varepsilon^{-1}(\delta^{-1} - 1)$, we get the individual dynamics $x^{\text{cur}} \mapsto \delta(x^{\text{orig}} - \varepsilon\theta) + (1-\delta)x^{\text{cur}}$, which in turn yields the mean map $m(\theta, \mu) = \delta(\mu_{\text{orig}} - \varepsilon\theta) + (1-\delta)\mu$. The point loss for this experiment is ridge-regularized cross-entropy.

The results are shown in Figure 3. DFO, PGD, and SPGD are all able to eventually find θ_{OPT}, but by simulating the long-term change in the distribution, SPGD is able to find this optimum in only 6 deployments. PGD requires long waits for the mean to settle in order to converge (leading to the flat regions in the training curve), and DFO requires deploying highly perturbed models to overcome the noise in the mean estimation.

5.5 Low-Dimensional Score

Finally, we test SPGD's performance in the setting described in Section 3.1 where the distribution dynamics are constrained by a low-dimensional bottleneck. The point loss is $\ell(z; \theta) = -z^\top\theta + \frac{\lambda}{2}\|\theta\|^2$, the score function is $s(\theta, \mu) = \theta^\top\mu$, and the stateful mean map is given by $m(\theta, \mu) = \bar{m}(s(\theta, \mu)) = (1 - \theta^\top\mu)\mu_0$ for some fixed μ_0. Under some restrictions on the model space Θ and the parameter μ_0, there exists a long-term distribution $\mu^*(\theta)$. See Appendix C for a derivation.

Refer to Figure 4. BSPGD (Bottleneck SPGD) refers to SPGD where we account for the one-dimensional bottleneck in the dynamics. Both SPGD and BSPGD are able to find θ_{OPT}, but by adapting the method to the one-dimensional score, BSPGD converges faster.

6 CONCLUSION

We considered the stateful performative setting and introduced Stateful PerfGD to optimize the long-term performative risk. We proved a convergence result for our method, and we verified empirically that our method is able to adapt to complicated stateful performative dynamics and find θ_{OPT}, whereas existing methods not tailored to this situation prove ineffective.

While our work does require parametric data assumptions, optimizing the performative loss for a fully general distribution map $\mathcal{D}$ is intractible. The parametric framework still provides a great deal of modeling flexibility, leaving the entire toolkit of parametric statistics

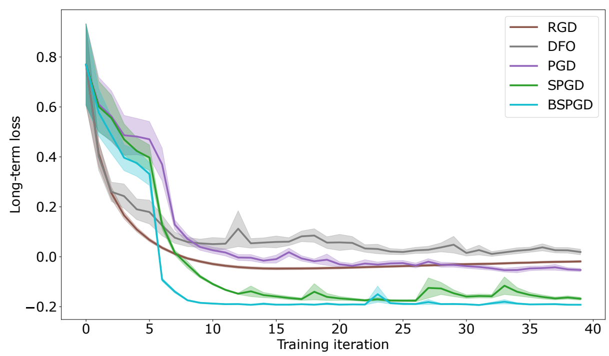

Figure 4: Performance of each method when the distribution shift depends only on a low-dimensional statistic. BSPGD denotes SPGD adapted to this setting. While vanilla SPGD outperforms both DFO and PGD, by taking into account the low-dimensional dependence, we can get even faster and more accurate convergence.

available to the user. The assumption of a fixed long-term distribution may also appear restrictive, but for many types of performative effects—such as strategic behavior on the part of the modeled population—this assumption will indeed hold, as the agents will have no incentive to change their behavior in the face of a fixed model once their desired outcome has been reached.

6.1 Societal Impact

The goal of optimization in the performative setting is to minimize the test loss. This is accomplished by choosing a model which is both accurate and induces a favorable data distribution, where "favorable" is measured only with respect to the model's goal. When the population in question consists of people, this amounts to trying to induce these people to behave in a way which makes them easy to classify, which may not align with behaviors that benefit these people the most. Indeed, it has been observed that in some cases, such a procedure can maximize a certain measure of negative externality (Jagadeesan et al., 2021). However, manipulation of the data distribution also has the capability to produce the opposite effect, i.e., inducing a data distribution which is advantageous both for the modeled population and the modeler. The distribution induced by the optimal model should also be studied to address these concerns.

6.2 Future Work

There are a number of interesting directions for future work. While minimizing the long-term performative risk is a sensible goal, other goals can also be considered—for instance, we can attempt to minimize the total loss incurred over the whole time horizon. In its current form, the problem is equivalent to a determinstic and highly structured Markov decision process, but relaxing some of the assumptions on the underlying MDP is of interest for improving the practical efficacy of this setting, and offers the potential for connections with reinforcement learning. Lastly, our current method works in the batch setting where we have enough samples to accurately estimate population-level quantities. Developing methods that can work in a stochastic/limited sample regime is also of interest.

Acknowledgements

J.Z. is supported by NSF CAREER 1942926 and grants from the Silicon Valley Foundation and the Chan–Zuckerberg Initiative. L.Y. is supported by the Scientific Discovery through Advanced Computing program, and by the National Science Foundation DMS-1818449. Z.I. is supported by a Stanford Interdisciplinary Graduate Fellowship.

References

Anderson, T. W. (1955). The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proceedings of the American Mathematical Society*, 6(2):170–176.

Brown, G., Hod, S., and Kalemaj, I. (2020). Performative prediction in a stateful world. *arXiv*.

Dong, J., Roth, A., Schutzman, Z., Waggoner, B., and Wu, Z. S. (2018). Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*.

Drusvyatskiy, D. and Xiao, L. (2020). Stochastic optimization with decision-dependent distributions. *arXiv*.

E, W. (2011). Principles of multiscale modeling. *Cambridge University Press*.

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. (2005). Online convex optimization in the bandit setting: Gradient descent without a gradient. *SODA*.

Ghalme, G., Nair, V., Eilat, I., Talgam-Cohen, I., and Rosenfeld, N. (2021). Strategic classification in the dark. *ICML*.

Hardt, M., Megiddo, N., Papadimitriou, C., and Wootters, M. (2016). Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Izzo, Z., Ying, L., and Zou, J. (2021). How to Learn when Data Reacts to Your Model: Performative Gradient Descent. *ICML*.

Jagadeesan, M., Mendler-Dünner, C., and Hardt, M. (2021). Alternative microfoundations for strategic classification. *ICML*.

Kayongo, P., Sun, G., Hartline, J., and Hullman, J. (2021). Visualization equilibrium. *IEEE Transactions on Visualization and Computer Graphics*.

Koh, P. W., Sagawa, S., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., et al. (2021). Wilds: A benchmark of in-the-wild distribution shifts. *ICML*.

Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. (2016). Gradient descent only converges to minimizers. *COLT*.

Levanon, S. and Rosenfeld, N. (2021). Strategic classification made practical. *ICML*.

Li, Q. and Wai, H.-T. (2021). State dependent performative prediction with stochastic approximation. *arXiv*.

Mendler-Dünner, C., Perdomo, J., Zrnic, T., and Hardt, M. (2020). Stochastic optimization for performative prediction. *NeurIPS*.

Miller, J., Perdomo, J. C., and Zrnic, T. (2021). Outside the Echo Chamber: Optimizing the Performative Risk. *ICML*.

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., and Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530.

Perdomo, J. C., Zrnic, T., Mendler-Dünner, C., and Hardt, M. (2020). Performative Prediction. *ICML*.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (2009). Dataset shift in machine learning. *The MIT Press*.

Sankar, A., Spielman, D. A., and Teng, S.-H. (2006). Smoothed analysis of the condition numbers and growth factors of matrices. *SIAM Journal on Matrix Analysis and Applications*, 28(2):446–476.

Storkey, A. (2009). When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*.

Sundaram, R., Vullikanti, A., Xu, H., and Yao, F. (2021). Pac-learning for strategic classification. *ICML*.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Wedin, P.-Å. (1973). Perturbation theory for pseudo-inverses. *BIT Numerical Mathematics*, 13(2):217–232.

Zhu, L., Peng, X., and Liu, H. (2019). Rank-one perturbation bounds for singular values of arbitrary matrices. *Journal of Inequalities and Applications*, 2019(1):1–10.

Zrnic, T., Mazumdar, E., Sastry, S. S., and Jordan, M. I. (2021). Who leads and who follows in strategic classification? *arXiv*.

A DERIVATION OF STATEFUL PERFGD

The long-term performative loss is given by

$$\mathcal{L}^*(\theta_t) = \int \ell(z;\theta_t) p(z;\mu^*(\theta_t))\, dz.$$

Its gradient is therefore given by

$$\nabla \mathcal{L}^*(\theta_t) = \int \nabla_\theta \ell(z;\theta_t)\, p(z;\mu^*(\theta_t))\, dz + \int \ell(z;\theta_t)\, \frac{d\mu^*}{d\theta}^\top \nabla_\mu p(z;\mu^*(\theta_t))\, dz.$$

The general form of our gradient estimate arises by substituting μ_t for $\mu^*(\theta_t)$ and $\widehat{\frac{dm^k}{d\theta}}$ for $\frac{d\mu^*}{d\theta}$.

The derivation for Algorithm 2 is as follows. For each time t, let $\psi_t = [\theta_t^\top, \mu_{t-1}^\top]^\top$, and define $m(\psi_t) = m(\theta_t, \mu_{t-1}) = \mu_t$. By Taylor's theorem, we have

$$m(\psi_s) - m(\psi_t) \approx \left.\frac{dm}{d\psi}\right|_{\psi_t} (\psi_s - \psi_t). \tag{3}$$

then we can vectorize equation (3) and obtain

$$\Delta\mu \approx \left.\frac{dm}{d\psi}\right|_{\psi_t} \Delta\psi \quad \Longrightarrow \quad \left.\frac{dm}{d\psi}\right|_{\psi_t} \approx (\Delta\mu)(\Delta\psi)^\dagger.$$

The expression for $\widehat{\frac{dm^{(k)}}{d\theta}}$ arises as follows. Observe that

$$\begin{aligned} \frac{d}{d\theta} m^{(k)}(\theta,\mu) &= \frac{d}{d\theta}[m(\theta, m^{(k-1)}(\theta,\mu))] \\ &= \partial_1 m(\theta, m^{(k-1)}(\theta,\mu)) \\ &+ \partial_2 m(\theta, m^{(k-1)}(\theta,\mu)) \cdot \frac{d}{d\theta} m^{(k-1)}(\theta,\mu). \end{aligned} \tag{4}$$

Since $m^{(k-1)}(\theta,\mu)$ is unknown, as are the derivatives $\partial_i m$ and $\frac{d}{d\theta} m^{(k-1)}(\theta,\mu)$, we simply substitute our "best guess" for each one. That is, we substitute μ for $m^{(k-1)}(\theta,\mu)$, $\widehat{\partial_i m}$ for $\partial_i m$, and $\widehat{\frac{dm^{(k-1)}}{d\theta}}$ for $\frac{dm^{(k-1)}}{d\theta}$. Thus we have

$$\widehat{\frac{dm^{(k)}}{d\theta}} = \partial_1 \widehat{m(\theta,\mu)} + \partial_2 \widehat{m(\theta,\mu)} \cdot \widehat{\frac{dm^{(k-1)}}{d\theta}},$$

with the base case $\widehat{\frac{dm^{(0)}}{d\theta}} = 0$ (the 0 matrix). Let $\widehat{\partial_i m} = \partial_i \widehat{m(\theta,\mu)}$. It can easily be shown via induction that

$$\widehat{\frac{dm^{(k)}}{d\theta}} = (I + \widehat{\partial_2 m} + (\widehat{\partial_2 m})^2 + \cdots + (\widehat{\partial_2 m})^{k-1})(\widehat{\partial_1 m}).$$

Assuming that $\|\widehat{\partial_2 m}\| < 1$ (which we expect to hold since $\|\partial_2 m\| < 1$), taking $k \to \infty$ yields

$$\lim_{k\to\infty} \widehat{\frac{dm^{(k)}}{d\theta}} = (I - \widehat{\partial_2 m})^{-1}(\widehat{\partial_1 m})$$

which is precisely the expression in (2).

B PROOFS FOR §4

B.1 Properties of the Gaussian distribution

In the proofs which follow, we make use of several key properties of the Gaussian distribution. Some of the well-known facts we state without proof.

Lemma 6. *Let $p(z, \mu)$ be the probability density function of a $\mathcal{N}(\mu, \Sigma)$ random variable, where Σ is a fixed covariance matrix. Then we have $\int \|\nabla_\mu p(z, \mu)\| \, dz \le \|\Sigma^{-1}\| \sqrt{d}$.*

Proof. We have

$$\begin{aligned}\int \|\nabla_\mu p(z, \mu)\| \, dz &= \int \|\Sigma^{-1}(z - \mu)\| p(z, \mu) \, dz \\ &\le \|\Sigma^{-1/2}\| \int \|\Sigma^{-1/2}(z - \mu)\| p(z, \mu) \, dz \\ &= \|\Sigma^{-1/2}\| \mathbb{E}_{z \sim \mathcal{N}(0, I_d)}[\|z\|] \qquad (5) \\ &\le \|\Sigma^{-1/2}\| \sqrt{d}. \qquad (6)\end{aligned}$$

Here (5) holds because $z \sim \mathcal{N}(\mu, \Sigma) \Rightarrow \Sigma^{-1/2}(z - \mu) \sim \mathcal{N}(0, I_d)$ and (6) holds by the well-known inequality $\mathbb{E}_{z \sim \mathcal{N}(0, I_d)}[\|z\|] \le \sqrt{d}$. ☐

Lemma 7. *Let $\sigma_0^2 = \|\Sigma\|$. Then we have*

$$\mathbb{P}_{z \sim \mathcal{N}(\mu, \Sigma)}(\|z - \mu\| \ge r + \sigma_0 \sqrt{d}) \le c_1 \exp\{-c_2 r^2 / \sigma_0^2\},$$

where c_1 is a constant which can depend on d and Σ, and c_2 is a universal constant.

Lemma 8. *Let $g \sim \mathcal{N}(0, \sigma^2 I) \in \mathbb{R}^d$. Then $\|g\| \le \sigma(\sqrt{d} + c\sqrt{\log \gamma^{-1}})$ with probability at least $1 - \gamma$ for some universal constant c. By a union bound, this means that with probability at least $1 - \gamma$, $\|g_t\| \le \sigma(\sqrt{d} + c\sqrt{\log \frac{T}{\gamma}}) = \widetilde{\mathcal{O}}(\sigma)$ for all $1 \le t \le T$.*

Lemma 9 (Anderson (1955))**.** *Suppose X and G are independent and $G \sim \mathcal{N}(0, \Sigma)$. Then for any $s > 0$, we have*

$$\mathbb{P}(\|X + G\| \le s) \le \mathbb{P}(\|G\| \le s).$$

Lemma 10. *Let $g \sim \mathcal{N}(0, I_n)$, and let $A \in \mathbb{R}^{n \times n}$ have singular values $s_1 \ge \cdots \ge s_n$. Suppose that $s_k \ge c$. Then $\mathbb{P}(\|Ag\| \le c\sqrt{k} - t) \le 2\exp(-c' \frac{t^2}{c^2})$ for some universal constant c'.*

Proof. Let $A = \sum_{i=1}^n s_i u_i v_i^\top$ be the SVD of A. Define $\tilde{g}_i = v_i^\top g$. Since the v_i form an orthonormal basis, we have $\tilde{g}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. Next, observe that

$$\begin{aligned}\|Ag\| &= \|\sum_{i=1^n} s_i (v_i^\top g) u_i\| \\ &= \|(s_1 \tilde{g}_1, \ldots, s_n \tilde{g}_n)\| \\ &\ge \|c(\tilde{g}_1, \ldots, \tilde{g}_k)\| \\ &= c\|\tilde{g}\|\end{aligned}$$

where $\tilde{g} = (\tilde{g}_1, \ldots, \tilde{g}_k)^\top \sim \mathcal{N}(0, I_k)$. The result then follows directly from (Vershynin, 2018), Theorem 3.1.1. ☐

Lemma 11. *Let $A \in \mathbb{R}^{n \times n}$ and let $s_1 \ge \cdots \ge s_n$ be the singular values of A. Suppose that $\|A\|_F \le cn$ and $\prod_{i=1}^n s_i \ge \beta^n$ for some $\beta > 0$. If $c_1\sqrt{n} \le k \le c_2\sqrt{n}$ for some universal constants c_1 and c_2 then there exists $n_0 = \mathcal{O}((\log \frac{c}{\beta})^2)$ such that for all $n \ge n_0$, $s_k \ge \frac{\beta}{2}$.*

Proof. Let k be arbitrary and suppose that $s_k < \frac{\beta}{2}$. Then we have

$$\beta^n \le \prod_{i=1}^n s_i \le \left(\prod_{i=1}^k s_i\right) \left(\frac{\beta}{2}\right)^{n-k} \quad \Longrightarrow \quad \prod_{i=1}^k s_i \ge 2^n \left(\frac{\beta}{2}\right)^k.$$

On the other hand, we have

$$\sum_{i=1}^k s_i \le \sum_{i=1}^n s_i = \|A\|_* \le \sqrt{n}\|A\|_F \le cn^{3/2}.$$

A simple Lagrange multiplier argument implies that $\max \prod_{i=1}^k s_i$ s.t. $\sum_{i=1}^k s_i \le C$ is $(C/k)^k$. Plugging in $C = cn^{3/2}$, we arrive at the inequality

$$2^n \left(\frac{\beta}{2}\right)^k \le \prod_{i=1}^k s_i \le \left(\frac{cn^{3/2}}{k}\right)^k \implies 2^n \le \left(\frac{2cn^{3/2}}{\beta k}\right)^k \implies n \le k\left(\log_2 \frac{2c}{\beta} + \log_2 \frac{n^{3/2}}{k}\right). \tag{7}$$

Now because $c_1\sqrt{n} \le k \le c_2\sqrt{n}$, (7) implies that

$$n^{1/2} \le c_2(\log_2 \frac{2c}{\beta} + \log_2 \frac{n}{c_1}) \le c_2 \log_2 \frac{2c}{\beta} + c_3 n^{1/4} \tag{8}$$

for some universal constant c_3. Now inequality (8) is quadratic in $n^{1/4}$, so applying the quadratic formula and simplifying, we see that it can only hold when $n \le n_0$ for some $n_0 = \mathcal{O}((\log \frac{c}{\beta})^2)$. This completes the proof. □

B.2 Useful Properties of m and μ^*

We will make use of the fact that $m^{(k+l)}(\theta, \mu) = m^{(k)}(\theta, m^{(l)}(\theta, \mu))$ for any $k, l \ge 0$. This is a simple consequence of the fact that $m^{(k)}(\theta, \mu)$ is the distribution parameters after k deployments of θ starting from μ, and deploying θ for l steps followed by k more deployments is the same as deploying θ for $k + l$ steps. It can also be shown rigorously by a simple double inductive argument.

Lemma 12. *For any $k \ge 0$, we have $\|\partial_2 m^{(k)}(\theta, \mu)\| \le \delta^k$. In particular, since $0 < \delta < 1$, we have $\|\partial_2 m^{(k)}(\theta, \mu)\| \le \delta$ for all $k \ge 1$.*

Proof. The claim is trivially true for $k = 0$. Inducting on k, we have:

$$\begin{aligned}
\|\partial_2 m^{(k+1)}(\theta, \mu)\| &= \|\frac{d}{d\mu} m(\theta, m^{(k)}(\theta, \mu))\| \\
&\le \|\partial_2 m(\theta, m^{(k)}(\theta, \mu))\| \cdot \|\partial_2 m^{(k)}(\theta, \mu)\| \\
&\le \delta \cdot \delta^k.
\end{aligned}$$

The above makes use of Assumption 2 and the inductive hypothesis. This completes the proof. □

Lemma 13. *For any $k \ge 0$, we have $\|\partial_1 m^{(k)}(\theta, \mu)\| \le \frac{B(1-\delta^k)}{1-\delta}$. In particular, since $0 < \delta < 1$, we have $\|\partial_1 m^{(k)}(\theta, \mu)\| \le \frac{B}{1-\delta}$ for all $k \ge 0$.*

Proof. The claim is trivially true for $k = 0$. Inducting on k, we have:

$$\begin{aligned}
\|\partial_1 m^{(k+1)}(\theta, \mu)\| &= \|\frac{d}{d\theta} m(\theta, m^{(k)}(\theta, \mu))\| \\
&\le \|\partial_1 m(\theta, m^{(k)}(\theta, \mu))\| + \|\partial_2 m(\theta, m^{(k)}(\theta, \mu))\|\|\partial_1 m^{(k)}(\theta, \mu)\| \\
&\le B + \delta \frac{B(1-\delta^k)}{1-\delta} \\
&= \frac{B(1-\delta^{k+1})}{1-\delta}.
\end{aligned}$$

The above uses Assumptions 1 and 2 and the inductive hypothesis. This completes the proof. □

Lemma 14. *There exists a function $\mu^*(\theta)$ such that $\lim_{k\to\infty} m^{(k)}(\theta, \mu) = \mu^*(\theta)$, independent of the starting parameters μ. Furthermore, for any θ, μ we have*

$$\|m^{(k)}(\theta, \mu) - \mu^*(\theta)\| \le D\delta^k.$$

Proof. Let μ be arbitrary and consider the sequence $m^{(k)}(\theta, \mu)$. We claim that this is a Cauchy sequence. WLOG let $k \le l$. Then by Lemma 12, we have

$$\begin{aligned}
\|m^{(k)}(\theta, \mu) - m^{(l)}(\theta, \mu)\| &= \|m^{(k)}(\theta, \mu) - m^{(k)}(\theta, m^{(l-k)}(\theta, \mu))\| \\
&\le \delta^k \|\mu - m^{(l-k)}(\theta, \mu)\| \\
&\le D\delta^k.
\end{aligned}$$

Since $0 < \delta < 1$, the sequence is Cauchy and therefore has a limit for any μ. Furthermore, again by Lemma 12, we have

$$\|m^{(k)}(\theta, \mu) - m^{(k)}(\theta, \mu')\| \leq \delta^k \|\mu - \mu'\| \leq D\delta^k,$$

which implies that the limit of these Cauchy sequences is independent of μ. We can thus set $\mu^*(\theta) = \lim_{k\to\infty} m^{(k)}(\theta, \mu)$ for any μ, and the above argument implies that μ^* is well-defined. It is also easy to see from this logic that $\mu^*(\theta)$ must be a fixed point of $m(\theta, \cdot)$.

Since $m^{(0)}(\theta, \mu) = \mu$ and $\|\mu - \mu^*(\theta)\| \leq D$ by definition of D, the claim holds for $k = 0$. We now induct and suppose the claim is true for arbitrary k. Then we have

$$\begin{aligned}\|m^{(k+1)}(\theta, \mu) - \mu^*(\theta)\| &= \|m(\theta, m^{(k)}(\theta, \mu)) - m(\theta, \mu^*(\theta))\| && (9)\\ &\leq \delta\|m^{(k)}(\theta, \mu) - \mu^*(\theta)\| && (10)\\ &\leq \delta \cdot D\delta^k.\end{aligned}$$

Here (9) holds by the recursive definition of $m^{(k+1)}$ and the fact that $\mu^*(\theta)$ is a fixed point of $m(\theta, \cdot)$, and (10) holds by Assumption 2. This completes the proof. ☐

Lemma 15. *The long-term parameters $\mu^*(\theta)$ are $\frac{B}{1-\delta}$-Lipschitz in θ, and therefore $\frac{d\mu^*}{d\theta}$ exists and we have $\|\frac{d\mu^*}{d\theta}\| \leq \frac{B}{1-\delta}$.*

Proof. By Lemma 13, $m^{(k)}$ are uniformly Lipschitz in θ with Lipschitz constant $B/(1-\delta)$. Since μ^* is the limit of Lipschitz functions with Lipschitz constants uniformly bounded by $B/(1-\delta)$, the result follows. ☐

Lemma 16. *Let $c = CD(1 + \frac{B}{1-\delta})$ and $B' = \frac{B}{1-\delta}$. Then $\|\frac{dm^{(k)}}{d\theta} - \frac{d\mu^*}{d\theta}\| \leq ck\delta^{k-1} + B'\delta^k = \mathcal{O}(k\delta^k)$. In particular, $\lim_{k\to\infty} \|\frac{dm^{(k)}}{d\theta} - \frac{d\mu^*}{d\theta}\| = 0$.*

Proof. In what follows, we will occasionally drop the dependence of $m^{(k)}(\theta, \mu)$ on θ and μ and the dependence of $\mu^*(\theta)$ on θ when this dependence is clear from context.

Since $\frac{dm^{(0)}}{d\theta} = 0$ and $\|\frac{d\mu^*}{d\theta}\| \leq B'$ by Lemma 15, the claim holds for $k = 0$. We induct:

$$\begin{aligned}\|\frac{dm^{(k+1)}}{d\theta} - \frac{d\mu^*}{d\theta}\| &= \|\frac{d}{d\theta}m(\theta, m^{(k)}(\theta, \mu)) - \frac{d}{d\theta}m(\theta, \mu^*(\theta))\|\\ &= \|\partial_1 m(\theta, m^{(k)}(\theta, \mu)) + \partial_2 m(\theta, m^{(k)}(\theta, \mu))\frac{dm^{(k)}}{d\theta} - \partial_1 m(\theta, \mu^*(\theta)) - \partial_2 m(\theta, \mu^*(\theta))\frac{d\mu^*}{d\theta}\|\\ &\leq \|\partial_1 m(\theta, m^{(k)}) - \partial_1 m(\theta, \mu^*)\| + \|\partial_2 m(\theta, m^{(k)}) - \partial_2 m(\theta, \mu^*)\|\|\frac{dm^{(k)}}{d\theta}\|\\ &+ \|\partial_2 m(\theta, \mu^*)\|\|\frac{dm^{(k)}}{d\theta} - \frac{d\mu^*}{d\theta}\|\\ &\leq C\|m^{(k)} - \mu^*\| + C\|m^{(k)} - \mu^*\|\frac{B}{1-\delta} + \delta\|\frac{dm^{(k)}}{d\theta} - \frac{d\mu^*}{d\theta}\|\\ &\leq CD(1 + \frac{B}{1-\delta})\delta^k + \delta\|\frac{dm^{(k)}}{d\theta} - \frac{d\mu^*}{d\theta}\|\\ &\leq c\delta^k + \delta(ck\delta^{k-1} + B'\delta^k)\\ &= c(k+1)\delta^k + B'\delta^{k+1}.\end{aligned}$$

The above makes use of Lemma 14 and Assumption 5. The second part of the lemma then holds since $0 < \delta < 1$. ☐

Lemma 17. *The norm of the gradient of the long-term performative loss is bounded by a constant:*

$$\|\nabla \mathcal{L}^*(\theta)\| \leq G = \mathcal{O}\left(\frac{\ell_{\max} B \|\Sigma^{-1/2}\| \sqrt{d}}{1-\delta}\right)$$

for all θ.

Proof. By Assumption 4 and Lemma 15, we have

$$\begin{aligned}\|\nabla \mathcal{L}^*(\theta)\| &\leq \int \|\nabla_\theta \ell(z;\theta)\|\, p(z;\mu^*(\theta))\,dz + \int |\ell(z;\theta)|\, \|\frac{d\mu^*}{d\theta}^\top\|\, \|\nabla_\mu p(z;\mu^*(\theta))\|\,dz \\ &\leq \ell_{\max} + \ell_{\max}\frac{B}{1-\delta}\int \|\nabla_\mu p(z,\mu^*(\theta))\|\,dz \\ &\leq G = \mathcal{O}\left(\frac{\ell_{\max} B\|\Sigma^{-1/2}\|\sqrt{d}}{1-\delta}\right).\end{aligned}$$

The last line follows from Lemma 6. ☐

We remark that, by a proof similar to the preceding, the bound on $\|\nabla^2\mathcal{L}^*(\theta)\|$ in Assumption 5 can be replaced with a bound on the second derivatives of ℓ. This constant will not appear in the leading order terms of our final bounds, so we opt for the simpler route of just assuming a priori that $\mathcal{L}^*$ has a bounded Hessian.

B.3 Approximation Results

Throughout, we will assume that $T = \mathcal{O}(\varepsilon^{-c})$ for some universal constant $c > 0$. (We can always stop the optimization procedure early if T is larger than this.) This implies that $\log T = o(\eta^{-c'})$, $o(\sigma^{-c})$ for any positive constant c'.

In our bounds, we will keep track of terms which are of leading order as η, σ, ε, $T^{-1} \to 0$. Given our eventual choices of η and σ, we will always have $\varepsilon = o(\eta)$, $\eta = o(\sigma)$, and $\sigma = o(1)$. We will also track the problem-dependent constants (e.g., B, C, D from the assumptions, the dimension d, etc.) which form coefficients for these leading order terms, but we still consider these as constants and therefore drop terms which are high order in η, σ, ε, T^{-1} but with worse dependence on the aforementioned constants. We also remark that we have not attempted to optimize our bounds with respect to these constants, and the dependence on them is likely not tight. Lastly, since the constant G defined in Lemma 17 appears frequently, we will make use of it rather than repeatedly writing $\frac{\ell_{\max}B\|\Sigma^{-1/2}\|\sqrt{d}}{1-\delta}$, but it should be noted that G can in fact be replaced by constants which exist a priori by the assumptions.

The overall structure of these proofs is inductive in nature. That is, we assume some conditions on the optimization trajectory so far—namely, bounds on the errors of various estimators—, and show that these properties continue to hold at the next step of the optimization.

We begin by showing that, after an initialization or "burn-in" period, the observed population means will be close to their equilibrium values. (In practice, the initialization can be quite short.) For ease of notation, we will always denote the θ update steps as $\hat{\nabla}\mathcal{L}_t^* + g_t$, though for the initialization phase we will just take $\hat{\nabla}\mathcal{L}_t^* = 0$. (That is, we initialize by updating θ_t by random Gaussian perturbations.)

Lemma 18. *Suppose that $t \geq \log\frac{1}{\eta}$, $\sigma = o(1/\sqrt{\log\frac{T}{\gamma}})$, and $\|\hat{\nabla}\mathcal{L}_s^*\| \leq cG$ for each $s < t$. Then we have*

$$\|\mu_t - \mu^*(\theta_t)\| = \mathcal{O}(BG(\log\frac{1}{\eta})^2\eta),$$

with probability at least $1-\gamma$ simultaneously for all t.

Proof. We claim that

$$\|\mu_t - \mu^*(\theta_t)\| \leq \delta B\sum_{l=1}^{k-1}\|\theta_{t-l} - \theta_t\| + \|m^{(k)}(\theta_t,\mu_{t-k}) - \mu^*(\theta_t)\| \tag{11}$$

for any $1 \leq k \leq t$. For $k = 1$, (11) is just the statement $\|\mu_t - \mu^*(\theta_t)\| \leq \|m^{(1)}(\theta_t,\mu_{t-1}) - \mu^*(\theta_t)\|$, which is true since $\mu_t = m^{(1)}(\theta_t, \mu_{t-1})$ and thus the LHS and RHS are equal. Now we induct on k:

$$\begin{aligned}
\|m^{(k)}(\theta_t, \mu_{t-k}) - \mu^*(\theta_t)\| &= \|m^{(k)}(\theta_t, m(\theta_{t-k}, \mu_{t-k-1})) - \mu^*(\theta_t)\| \\
&\le \|m^{(k)}(\theta_t, m(\theta_{t-k}, \mu_{t-k-1})) - m^{(k)}(\theta_t, m(\theta_t, \mu_{t-k-1}))\| \quad (12)\\
&+ \|m^{(k)}(\theta_t, m(\theta_t, \mu_{t-k-1})) - \mu^*(\theta_t)\| \\
&\le \delta\|m(\theta_{t-k}, \mu_{t-k-1}) - m(\theta_t, \mu_{t-k-1})\| + \|m^{(k)}(\theta_t, m(\theta_t, \mu_{t-k-1})) - \mu^*(\theta_t)\| \quad (13)\\
&\le \delta B\|\theta_{t-k} - \theta_t\| + \|m^{(k)}(\theta_t, m(\theta_t, \mu_{t-k-1})) - \mu^*(\theta_t)\|. \quad (14)
\end{aligned}$$

Here (13) follows from Lemma 12 and (14) uses Assumption 1. If we use the fact that $m^{(k+1)}(\theta, \mu_{t-k-1}) = m^{(k)}(\theta_t, m(\theta_t, \mu_{t-k-1}))$ (this is simply unrolling the recursive definition for $m^{(k+1)}$ from the inside out instead of outside in) and plug (14) into the inductive hypothesis, we complete the induction and (11) holds for all $k \le t$.

Next, for $l \le k$, observe that

$$\begin{aligned}
\|\theta_{t-l} - \theta_t\| &= \eta\|\hat{\nabla}\mathcal{L}^*_{t_l} + g_{t-l} + \cdots + \hat{\nabla}\mathcal{L}^*_{t-1} + g_{t-1}\| \\
&\le \eta \sum_{i=1}^{l} \|\hat{\nabla}\mathcal{L}^*_{t-i}\| + \|g_{t-i}\| \\
&\le \eta \cdot l \cdot (cG + o(1)) \quad (15)\\
&\le c''Gk\eta. \quad (16)
\end{aligned}$$

Here (15) holds since we have assumed the the high-probability guarantee of Lemma 8 holds for all t. Plugging this inequality into (11), we have that

$$\begin{aligned}
\|\mu_t - \mu^*(\theta_t)\| &\le \delta B \cdot (k-1) \cdot c''Gk\eta + D\delta^k \\
&= \mathcal{O}(BGk^2\eta + D\delta^k)
\end{aligned}$$

where we have also used Lemma 14 to bound $\|m^{(k)}(\theta_t, \mu_{t-k}) - \mu^*(\theta_t)\|$. Setting $k = \log \frac{1}{\eta}$ (which is valid since $t \ge \log \frac{1}{\eta}$), we obtain

$$\|\mu_t - \mu^*(\theta_t)\| = \mathcal{O}(BG(\log \frac{1}{\eta})^2\eta) = \widetilde{\mathcal{O}}(\eta).$$

□

Lemma 2. *Suppose that $\|\hat{\nabla}\mathcal{L}^*_s\|$ are bounded by a constant for $s < t$. Let $\frac{dm}{d\psi}$ denote the true Jacobian of m with respect to its input, and let $\widehat{\frac{dm}{d\psi}} = [\widehat{\partial_1 m}, \widehat{\partial_2 m}]$ denote the estimator from Algorithm 2. Then for appropriate choices of η, σ, and H, we have*

$$\left\|\widehat{\frac{dm}{d\psi}} - \frac{dm}{d\psi}\right\| \le \widetilde{\mathcal{O}}\left(\frac{\eta}{\sigma} + \frac{\varepsilon}{\eta\sigma^2}\right) \equiv \mathbf{E}_m.$$

Proof. We begin by decomposing $\Delta\mu$ and $\Delta\psi$.

$$\begin{aligned}
\Delta\mu &= \begin{bmatrix}\hat{\mu}_{t-1} - \hat{\mu}_t & \cdots & \hat{\mu}_{t-H} - \hat{\mu}_t\end{bmatrix} \\
&= \underbrace{\begin{bmatrix}\mu_{t-1} - \mu_t & \cdots & \mu_{t-H} - \mu_t\end{bmatrix}}_{\overline{\Delta\mu}} + \underbrace{\begin{bmatrix}\varepsilon_{t-1} - \varepsilon_t & \cdots & \varepsilon_{t-H} - \varepsilon_t\end{bmatrix}}_{\text{err}_\mu}
\end{aligned}$$

$$\begin{aligned}
\Delta\psi &= \begin{bmatrix}\theta_{t-1} - \theta_t & \cdots & \theta_{t-H} - \theta_t \\ \hat{\mu}_{t-2} - \hat{\mu}_{t-1} & \cdots & \hat{\mu}_{t-H-1} - \hat{\mu}_{t-1}\end{bmatrix} \\
&= \underbrace{\begin{bmatrix}\theta_{t-1} - \theta_t & \cdots & \theta_{t-H} - \theta_t \\ \mu_{t-2} - \mu_{t-1} & \cdots & \mu_{t-H-1} - \mu_{t-1}\end{bmatrix}}_{\overline{\Delta\psi}} + \underbrace{\begin{bmatrix}0 & \cdots & 0 \\ \varepsilon_{t-2} - \varepsilon_{t-1} & \cdots & \varepsilon_{t-H-1} - \varepsilon_{t-1}\end{bmatrix}}_{\text{err}_\psi}
\end{aligned}$$

Using this expression, we can rewrite $\widehat{\frac{dm}{d\psi}}$:

$$\widehat{\frac{dm}{d\psi}} = (\overline{\Delta\mu})(\overline{\Delta\psi})^\dagger + (\mathrm{err}_\mu)(\overline{\Delta\psi})^\dagger + (\Delta\mu)[(\overline{\Delta\psi} + \mathrm{err}_\psi)^\dagger - (\overline{\Delta\psi})^\dagger].$$

This allows us to decompose the error of $\widehat{\frac{dm}{d\psi}}$ as

$$\|\frac{dm}{d\psi} - (\Delta\mu)(\Delta\psi)^\dagger\| \leq \|\frac{dm}{d\psi} - (\overline{\Delta\mu})(\overline{\Delta\psi})^\dagger\| + \|\mathrm{err}_\mu\|\|\overline{\Delta\psi}^\dagger\| + \underbrace{\|\Delta\mu\|\|(\overline{\Delta\psi} + \mathrm{err}_\psi)^\dagger - (\overline{\Delta\psi})^\dagger\|}_{(\star)}. \tag{17}$$

Before we begin, let us decompose $(\star)$ further. By (Wedin, 1973), if $\Delta\psi$ and $\overline{\Delta\psi}$ are both full rank, then

$$\|(\overline{\Delta\psi} + \mathrm{err}_\psi)^\dagger - (\overline{\Delta\psi})^\dagger\| \leq \sqrt{2}\|(\overline{\Delta\psi} + \mathrm{err}_\psi)^\dagger\|\|\overline{\Delta\psi}^\dagger\|\|\mathrm{err}_\psi\|. \tag{18}$$

By Weyl's inequality for singular values, we also have

$$\begin{aligned}\|(\overline{\Delta\psi} + \mathrm{err}_\psi)^\dagger\| &= \frac{1}{\sigma_{\min}(\overline{\Delta\psi} + \mathrm{err}_\psi)} \\ &\leq \frac{1}{\sigma_{\min}(\overline{\Delta\psi}) - \sigma_{\max}(\mathrm{err}_\psi)} \\ &= \frac{1}{\frac{1}{\|\overline{\Delta\psi}^\dagger\|} - \|\mathrm{err}_\psi\|} \\ &= \frac{\|\overline{\Delta\psi}^\dagger\|}{1 - \|\overline{\Delta\psi}^\dagger\|\|\mathrm{err}_\psi\|}. \end{aligned} \tag{19}$$

Combining (18) and (19) and substituting them into (17), we obtain

$$\|\frac{dm}{d\psi} - (\Delta\mu)(\Delta\psi)^\dagger\| \leq \underbrace{\|\frac{dm}{d\psi} - (\overline{\Delta\mu})(\overline{\Delta\psi})^\dagger\|}_{\text{(I)}} + \underbrace{\|\mathrm{err}_\mu\|\|\overline{\Delta\psi}^\dagger\|}_{\text{(II)}} + \sqrt{2}\underbrace{\|\Delta\mu\|\frac{\|\overline{\Delta\psi}^\dagger\|^2}{1 - \|\overline{\Delta\psi}^\dagger\|\|\mathrm{err}_\psi\|}\|\mathrm{err}_\psi\|}_{\text{(III)}}. \tag{20}$$

Let us address (I) first. Recall that $\frac{dm}{d\psi}$ refers to the Jacobian of m evaluated at (θ_t, μ_{t-1}). By Taylor's theorem, for any $1 \leq i \leq H$, we have

$$\mu_{t-i} - \mu_t = \frac{dm}{d\psi}(\psi_{t-i} - \psi_t) + \nabla^2 m(\xi_i)[\psi_{t-i} - \psi_t,\, \psi_{t-i} - \psi_t]. \tag{21}$$

Here $\psi_{t-i} = (\theta_{t-i}^\top, \mu_{t-1-i}^\top)^\top$ and $\nabla^2 m(\xi_i)$ denotes the tensor of second derivatives of m evaluated at some point ξ_i specified by Taylor's theorem. For each ξ_i, $\nabla^2 m(\xi_i)$ is a bilinear map from $\mathbb{R}^{2d} \times \mathbb{R}^{2d} \to \mathbb{R}^d$, and $\nabla^2 m(\xi_i)[\psi_{t-i} - \psi_t,\, \psi_{t-i} - \psi_t]$ denotes evaluation of this map at the inputs specified in the brackets. Define $e_i^{\text{Taylor}} = \nabla^2 m(\xi_i)[\psi_{t-i} - \psi_t,\, \psi_{t-i} - \psi_t]$. By (21), we have

$$\begin{aligned}\overline{\Delta\mu} &= \begin{bmatrix}\frac{dm}{d\psi}(\psi_{t-1} - \psi_t) + e_1^{\text{Taylor}} & \cdots & \frac{dm}{d\psi}(\psi_{t-H} - \psi_t) + e_H^{\text{Taylor}}\end{bmatrix} \\ &= \frac{dm}{d\psi}\begin{bmatrix}\psi_{t-1} - \psi_t & \cdots & \psi_{t-H} - \psi_t\end{bmatrix} + \begin{bmatrix}e_1^{\text{Taylor}} & \cdots & e_H^{\text{Taylor}}\end{bmatrix} \\ &= \frac{dm}{d\psi}\overline{\Delta\psi} + \underbrace{\begin{bmatrix}e_1^{\text{Taylor}} & \cdots & e_H^{\text{Taylor}}\end{bmatrix}}_{E^{\text{Taylor}}}. \end{aligned} \tag{22}$$

Assume for the moment that $\overline{\Delta\psi}$ has full rank; we will prove this later. Since we have chosen $H \geq 2d$ and $\overline{\Delta\psi} \in \mathbb{R}^{2d\times H}$, this implies that $\overline{\Delta\psi}$ has a right inverse. Combining this fact with (22), we have

$$\begin{aligned}
\text{(I)} &= \|\frac{dm}{d\psi} - (\overline{\Delta\mu})(\overline{\Delta\psi})^\dagger\| \\
&= \|\frac{dm}{d\psi} - (\frac{dm}{d\psi} + E^{\text{Taylor}}(\overline{\Delta\psi})^\dagger)\| \\
&\leq \|E^{\text{Taylor}}\|\|\overline{\Delta\psi}^\dagger\|. \qquad (23)
\end{aligned}$$

We now bound $\|E^{\text{Taylor}}\|$. Because the operator norm $\|\nabla^2 m\| \leq C$ by Assumption 5, we have

$$\|e_i^{\text{Taylor}}\| = \|\nabla^2 m(\xi_i)[\psi_{t-i} - \psi_t,\, \psi_{t-i} - \psi_t]\| \leq C\|\psi_{t-i} - \psi_t\|^2. \qquad (24)$$

It follows that

$$\begin{aligned}
\|E^{\text{Taylor}}\| &\leq \|E^{\text{Taylor}}\|_F \\
&= \sqrt{\sum_{i=1}^{H} \|e_i^{\text{Taylor}}\|^2} \\
&\leq \sqrt{\sum_{i=1}^{H} (C\|\psi_{t-i} - \psi_t\|^2)^2}. \qquad (25)
\end{aligned}$$

By definition of ψ_{t-i} and ψ_t, we have

$$\begin{aligned}
\|\psi_{t-i} - \psi_t\|^2 &= \left\| \begin{bmatrix} \theta_{t-i} - \theta_t \\ \mu_{t-1-i} - \mu_{t-1} \end{bmatrix} \right\|^2 \\
&= \|\theta_{t-i} - \theta_t\|^2 + \|\mu_{t-1-i} - \mu_{t-1}\|^2 \\
&\leq \mathcal{O}((GH\eta)^2) + (\|\mu_{t-1-i} - \mu^*(\theta_{t-1-i})\| + \|\mu^*(\theta_{t-1-i}) - \mu^*(\theta_{t-1})\| + \|\mu^*(\theta_{t-1}) - \mu_{t-1}\|)^2 \qquad (26) \\
&\leq \mathcal{O}(G^2H^2\eta^2) + (\mathcal{O}(BG(\log\frac{1}{\eta})^2\eta) + \frac{B}{1-\delta}\|\theta_{t-1-i} - \theta_{t-1}\| + \mathcal{O}(BG(\log\frac{1}{\eta})^2\eta))^2 \qquad (27) \\
&\leq \mathcal{O}(G^2H^2\eta^2) + [\mathcal{O}(BG(\log\frac{1}{\eta})^2\eta + \frac{B}{1-\delta}GH\eta)]^2 \qquad (28) \\
&= \mathcal{O}(B^2G^2(\log\frac{1}{\eta})^4\eta^2). \qquad (29)
\end{aligned}$$

Inequality (26) holds by the logic from (16), as well as by splitting $\|\mu_{t-1-i} - \mu_{t-1}\|$ with the triangle inequality. Inequality (27) holds by Lemmas 15 and 18. Finally, inequality (28) again holds by the logic from (16).

Plugging the result of (29) into (25), we have

$$\begin{aligned}
\|E^{\text{Taylor}}\| &\leq C\sqrt{\sum_{i=1}^{H} [\mathcal{O}(B^2G^2(\log\frac{1}{\eta})^4\eta^2)]^2} \\
&\leq C\sqrt{H \cdot [\mathcal{O}(B^2G^2(\log\frac{1}{\eta})^4\eta^2)]^2} \\
&= \mathcal{O}(C\sqrt{H}B^2G^2(\log\frac{1}{\eta})^4\eta^2) \\
&= \widetilde{\mathcal{O}}(\eta^2). \qquad (30)
\end{aligned}$$

Combining this with (23) yields that

$$(\mathrm{I}) = \mathcal{O}(C\sqrt{H}B^2G^2(\log\frac{1}{\eta})^4\eta^2)\|\overline{\Delta\psi}^\dagger\|. \tag{31}$$

Next, observe that since $\|\varepsilon_{t-i}\| \leq \boldsymbol{\varepsilon}$, we have

$$\|\mathrm{err}_\mu\| \leq \|\mathrm{err}_\mu\|_F = \mathcal{O}(\sqrt{H}\boldsymbol{\varepsilon}), \qquad \|\mathrm{err}_\psi\| \leq \|\mathrm{err}_\psi\|_F = \mathcal{O}(\sqrt{H}\boldsymbol{\varepsilon}). \tag{32}$$

We now turn our attention back to (III); in particular we will bound $\|\Delta\mu\|$. By the definition of $\overline{\Delta\mu}$, we have

$$\begin{aligned}
\|\Delta\mu\| &\leq \|\overline{\Delta\mu}\| + \|\mathrm{err}_\mu\| \\
&\leq \|\frac{dm}{d\psi}\|\|\overline{\Delta\psi}\| + \|E^{\mathrm{Taylor}}\| + \|\mathrm{err}_\mu\| \qquad (33)\\
&= \mathcal{O}\left(B\|\overline{\Delta\psi}\| + C\sqrt{H}B^2G^2(\log\frac{1}{\eta})^4\eta^2 + \sqrt{H}\boldsymbol{\varepsilon}\right). \qquad (34)
\end{aligned}$$

Here (33) follows from equation (22). Equation (34) follows from (30), (32), and Assumptions 1 and 2.

To bound $\|\overline{\Delta\psi}\|$, we make use of (29). We have

$$\begin{aligned}
\|\overline{\Delta\psi}\| &\leq \|\overline{\Delta\psi}\|_F \\
&= \sqrt{\sum_{i=1}^{H}\|\psi_{t-i} - \psi_t\|^2} \\
&\leq \sqrt{H \cdot \mathcal{O}(B^2G^2(\log\frac{1}{\eta})^4\eta^2)} \qquad (35)\\
&= \mathcal{O}(BG\sqrt{H}(\log\frac{1}{\eta})^2\eta) \\
&= \widetilde{\mathcal{O}}(\eta). \qquad (36)
\end{aligned}$$

Inequality (35) follows from (29). Plugging (36) into (34), we find that

$$\begin{aligned}
\|\Delta\mu\| &= \mathcal{O}(B^2G\sqrt{H}(\log\frac{1}{\eta})^2\eta + \sqrt{H}\boldsymbol{\varepsilon}) \\
\Rightarrow (\mathrm{III}) &= \mathcal{O}(B^2G\sqrt{H}(\log\frac{1}{\eta})^2\eta) \cdot \frac{\|\overline{\Delta\psi}^\dagger\|^2}{1 - \|\overline{\Delta\psi}^\dagger\|\boldsymbol{\varepsilon}} \cdot \mathcal{O}(\sqrt{H}\boldsymbol{\varepsilon}) \qquad (37)\\
&= \mathcal{O}(B^2GH(\log\frac{1}{\eta})^2\eta\boldsymbol{\varepsilon}\|\overline{\Delta\psi}^\dagger\|^2). \qquad (38)
\end{aligned}$$

Equation (37) uses the definition of (III) and the bound on $\|\Delta\mu\|$ (which we have reduced using that fact that $\boldsymbol{\varepsilon} = o(\eta)$), as well as inequality (32). Equation (38) holds under the assumption that $\|\overline{\Delta\psi}^\dagger\|\|\mathrm{err}_\psi\| \leq \frac{1}{2}$, which we will show holds given the final choice of η and σ.

Let us combine what we have shown so far to reduce the expression (20):

$$\|\frac{dm}{d\psi} - (\Delta\mu)(\Delta\psi)^\dagger\| = \mathcal{O}\left(C\sqrt{H}B^2G^2(\log\frac{1}{\eta})^4\eta^2\|\overline{\Delta\psi}^\dagger\| + \sqrt{H}\boldsymbol{\varepsilon}\|\overline{\Delta\psi}^\dagger\| + B^2GH(\log\frac{1}{\eta})^2\eta\boldsymbol{\varepsilon}\|\overline{\Delta\psi}^\dagger\|^2\right). \tag{39}$$

We have thus reduced the problem to showing that $\overline{\Delta\psi}$ has full rank and bounding $\|\overline{\Delta\psi}^\dagger\|$ from above. By

Lemma 19 below, $\|\overline{\Delta\psi}^\dagger\| = \mathcal{O}(\frac{1}{\eta\sigma})$. Plugging this into (39), we arrive at

$$\begin{aligned}\|\frac{dm}{d\psi} - (\Delta\mu)(\Delta\psi)^\dagger\| &= \mathcal{O}\left(C\sqrt{H}B^2G^2(\log\frac{1}{\eta})^4\eta^2\frac{1}{\eta\sigma} + \sqrt{H}\varepsilon\frac{1}{\eta\sigma} + B^2GH(\log\frac{1}{\eta})^2\eta\varepsilon\frac{1}{\eta^2\sigma^2}\right) \\ &= \mathcal{O}\left(C\sqrt{H}B^2G^2(\log\frac{1}{\eta})^4\frac{\eta}{\sigma} + B^2GH(\log\frac{1}{\eta})^2\frac{\varepsilon}{\eta\sigma^2}\right) \equiv \mathbf{E}_m.\end{aligned}$$

This completes the proof. ∎

Lemma 19. *Suppose that* $\|\hat{\nabla}\mathcal{L}^*(\theta_{t-i}) - \nabla\mathcal{L}^*(\theta_{t-i})\| \leq \mathbf{E}_\nabla$, *where*

$$\mathbf{E}_\nabla = \mathcal{O}\left(\frac{CB^3G^2\sqrt{Hd}\|\Sigma^{-1/2}\|}{(1-\delta)^2}(\log\frac{1}{\eta})^4\frac{\eta}{\sigma} + \frac{B^3GH\sqrt{d}\|\Sigma^{-1/2}\|}{(1-\delta)^2}(\log\frac{1}{\eta})^2\frac{\varepsilon}{\eta\sigma^2}\right) = \widetilde{\mathcal{O}}(\frac{\eta}{\sigma} + \frac{\varepsilon}{\eta\sigma^2}). \tag{40}$$

Furthermore, assume that

$$\sigma \geq 2c'\frac{BH^{3/2}\mathbf{E}_\nabla}{1-\delta} \qquad \text{and} \qquad \sigma = o(1)$$

where c' is an absolute constant to be specified later. Then with probability at least $1 - \frac{3\gamma}{T}$, $\overline{\Delta\psi}$ has full rank and $\|\overline{\Delta\psi}^\dagger\| = \mathcal{O}(\frac{1}{\eta\sigma})$ as long as $H = \Theta(\frac{B^8d^4}{\alpha^8(1-\delta)^8}(\log\frac{T}{\gamma})^4)$.

Proof. Define $e_{t-i} = \hat{\nabla}\mathcal{L}^*(\theta_{t-i}) - \nabla\mathcal{L}^*(\theta_{t-i})$ for $i = 1,\dots,H$, so $\|e_{t-i}\| \leq \mathbf{E}_\nabla$ for $i = 1,\dots,H$. (Note: If $t-i \leq \log\frac{1}{\eta}$ so that step $t-i$ was during the intialization phase, then $\hat{\nabla}\mathcal{L}^*(\theta_{t-i})$ and $\nabla\mathcal{L}^*(\theta_{t-i})$ are both replaced with 0 and the error is 0 for these steps. The logic that follows can trivially be extended to this case.)

Claim 1: We have

$$\theta_{t-H+i} = \bar{\theta}_{t-H+i} - \eta\underbrace{\sum_{j=0}^{i-1} g_{t-H+j}}_{G_{t-H+i}} - \eta\underbrace{\sum_{j=0}^{i-1} e_{t-H+j}}_{E_{t-H+i}} + \eta^2 S_{t-H+i},$$

where $\bar{\theta}_{t-H+i}$ is independent of all of the Gaussian perturbations g_{t-H+j}, $j = 0,\dots,H-1$, and S_{t-H+i} is defined recursively by

$$S_{t-H+i} = \sum_{j=0}^{i-1}\nabla^2\mathcal{L}^*(\zeta_{t-H+j})(G_{t-H+j} + E_{t-H+j} - \eta S_{t-H+j})$$

for some collection of vectors ζ_{t-H+j} in $\mathbb{R}^d$.

Proof of Claim 1: The claim holds trivially when $i = 0$, so we induct on i. We have:

$$\begin{aligned}\theta_{t-H+i+1} &= \theta_{t-H+i} - \eta(\nabla\mathcal{L}^*(\theta_{t-H+i}) + g_{t-H+i} + e_{t-H+i}) \\ &= \theta_{t-H+i} - \eta(g_{t-H+i} + e_{t-H+i}) - \eta\nabla\mathcal{L}^*(\bar{\theta}_{t-H+i} - \eta G_{t-H+i} - \eta E_{t-H+i} + \eta^2 S_{t-H+i}) \\ &= \theta_{t-H+i} - \eta(g_{t-H+i} + e_{t-H+i}) \\ &\quad - \eta(\nabla\mathcal{L}^*(\bar{\theta}_{t-H+i}) - \eta\nabla^2\mathcal{L}^*(\zeta_{t-H+i})(G_{t-H+i} + E_{t-H+i} - \eta S_{t-H+i})) \qquad (41) \\ &= \bar{\theta}_{t-H+i} - \eta\nabla\mathcal{L}^*(\bar{\theta}_{t-H+i}) - \eta(G_{t-H+i} + g_{t-H+i}) - \eta(E_{t-H+i} + e_{t-H+i}) \\ &\quad + \eta^2(S_{t-H+i} + \nabla^2\mathcal{L}^*(\zeta_{t-H+i})(G_{t-H+i} + E_{t-H+i} - \eta S_{t-H+i}).\end{aligned}$$

Equation (41) follows by Taylor expanding $\nabla\mathcal{L}^*$ about $\bar{\theta}_{t-H+i}$. We observe that

$$\begin{aligned}G_{t-H+i} + g_{t-H+i} &= \sum_{j=0}^{i-1} g_{t-H+j} + g_{t-H+i} \\ &= \sum_{j=0}^{i} g_{t-H+j} \\ &= G_{t-H+i+1}.\end{aligned}$$

Similarly, we have

$$E_{t-H+i} + e_{t-H+i} = E_{t-H+i+1}$$

$$S_{t-H+i} + \nabla^2 \mathcal{L}^*(\zeta_{t-H+i})(G_{t-H+i} + E_{t-H+i} - \eta S_{t-H+i}) = S_{t-H+i+1},$$

which completes the induction and proves Claim 1.

Before we proceed, we first bound $\|E_{t-H+i}\|$ and $\|S_{t-H+i}\|$. By the formula for E_{t-H+i} and the fact that $\|e_{t-H+j}\| \leq \mathbf{E}_\nabla$ for all j, it is clear that

$$\|E_{t-H+i}\| \leq H\mathbf{E}_\nabla$$

for all i. Furthermore, we have assumed that we are on the high-probability event that $\|g_t\| = \mathcal{O}(\sigma(\sqrt{d} + \sqrt{\log \frac{T}{\gamma}}))$ for all $1 \leq t \leq T$ (see Lemma 8). This implies that

$$\|G_{t-H+i}\| = \mathcal{O}(H\sigma(\sqrt{d} + \sqrt{\log \frac{T}{\gamma}})) = \widetilde{\mathcal{O}}(\sigma)$$

for all i. Thus we can choose $\mathbf{G}$ such that $\|G_{t-H+i}\| \leq \mathbf{G} \leq cH\sigma(\sqrt{d} + \sqrt{\log \frac{T}{\gamma}})$ for some universal constant c.

Finally, we claim that $\|S_{t-H+i}\| \leq \frac{(1+L\eta)^i - 1}{L\eta} \cdot (\mathbf{G} + H\mathbf{E}_\nabla)$ for all i. Since $S_{t-H} = 0$, the claim holds for $i = 0$. We induct:

$$\begin{aligned}
\|S_{t-H+i+1}\| &\leq \|S_{t-H+i}\| + \|\nabla^2 \mathcal{L}^*(\zeta_{t-H+i})\|(\|G_{t-H+i}\| + \|E_{t-H+i}\| + \eta\|S_{t-H+i}\|) \\
&\leq (1+L\eta)\|S_{t-H+i}\| + L(\mathbf{G} + H\mathbf{E}_\nabla) && (42) \\
&\leq \frac{(1+L\eta)^{i+1} - (1+L\eta)}{L\eta} \cdot (\mathbf{G} + H\mathbf{E}_\nabla) + (\mathbf{G} + H\mathbf{E}_\nabla) && (43) \\
&= \frac{(1+L\eta)^{i+1} - 1}{L\eta} \cdot (\mathbf{G} + H\mathbf{E}_\nabla). && (44)
\end{aligned}$$

This completes the induction. Since $i \leq H$, we have $\|S_{t-H+i}\| \leq \frac{(1+L\eta)^H - 1}{L\eta}(\mathbf{G} + H\mathbf{E}_\nabla)$. Since $\eta = o(1)$, by expanding, we see that $(1+L\eta)^H = 1 + \mathcal{O}(HL\eta)$. It follows that

$$\|S_{t-H+i}\| \leq \frac{1 + \mathcal{O}(HL\eta) - 1}{L\eta}(\mathbf{G} + H\mathbf{E}_\nabla) = \mathcal{O}(H(\mathbf{G} + H\mathbf{E}_\nabla)) = \mathcal{O}\left(H^2(\sigma(\sqrt{d} + \sqrt{\log \frac{T}{\gamma}}) + \mathbf{E}_\nabla)\right).$$

To summarize, we have:

$$\|E_{t-H+i}\| = \mathcal{O}(H\mathbf{E}_\nabla), \qquad \|G_{t-H+i}\| = \mathcal{O}\left(H\left(\sqrt{d} + \sqrt{\log \frac{T}{\gamma}}\right)\sigma\right), \tag{45}$$

$$\|S_{t-H+i}\| = \mathcal{O}\left(H^2\left(\sqrt{d} + \sqrt{\log \frac{T}{\gamma}}\right)\sigma + H^2\mathbf{E}_\nabla\right).$$

Claim 2: We have the decomposition

$$\mu_{t-H+i} = \bar{\mu}_{t-H+i} - \eta G^\mu_{t-H+i} - \eta E^\mu_{t-H+i} + \eta^2 S^\mu_{t-H+i},$$

where $\bar{\mu}_{t-H+i}$ is independent of all of the Gaussian perturbations g_{t-H+j} and the remaining terms in the expression are given by recursive definitions below.

Proof of Claim 2: The claim holds trivially when $i = -1$, so we induct on i. In what follows, to avoid notational clutter, we replace the index $t - H + i$ by i. Define $\bar{\psi}_i = (\bar{\theta}_i,\ \bar{\mu}_{i-1})$, and set $\partial_1 m_i = \partial_1 m(\bar{\psi}_i)$ and $\partial_2 m_i = \partial_2 m(\bar{\psi}_i)$. Lastly, define

$$U_i = \begin{bmatrix} G_i + E_i - \eta S_i \\ G^{\mu}_{i-1} + E^{\mu}_{i-1} - \eta S^{\mu}_{i-1} \end{bmatrix}.$$

We have:

$$\begin{aligned}
\mu_{i+1} &= m(\theta_{i+1},\ \mu_i) \\
&= m(\bar{\theta}_{i+1} - \eta G_{i+1} - \eta E_{i+1} + \eta^2 S_{i+1},\ \ \bar{\mu}_i - \eta G^{\mu}_i - \eta E^{\mu}_i + \eta^2 S^{\mu}_i) \\
&= m(\bar{\psi}_{i+1}) \\
&\quad - \eta \partial_1 m(\bar{\psi}_{i+1})(G_{i+1} + E_{i+1}) - \eta \partial_2 m(\bar{\psi}_{i+1})(G^{\mu}_i + E^{\mu}_i) \\
&\quad + \eta^2 \partial_1 m(\bar{\psi}_{i+1}) S_{i+1} + \eta^2 \partial_2 m(\bar{\psi}_{i+1}) S^{\mu}_i + \eta^2 \nabla^2 m(\omega_{i+1})[\begin{bmatrix} G_{i+1} + E_{i+1} - \eta S_{i+1} \\ G^{\mu}_i + E^{\mu}_i - \eta S^{\mu}_i \end{bmatrix}, \begin{bmatrix} G_{i+1} + E_{i+1} - \eta S_{i+1} \\ G^{\mu}_i + E^{\mu}_i - \eta S^{\mu}_i \end{bmatrix}]
\end{aligned} \tag{46}$$

$$\begin{aligned}
&= m(\bar{\psi}_{i+1}) - \eta((\partial_1 m_{i+1}) G_{i+1} + (\partial_2 m_{i+1}) G^{\mu}_i) - \eta((\partial_1 m_{i+1}) E_{i+1} + (\partial_2 m_{i+1}) E^{\mu}_{i+1}) \\
&\quad + \eta^2 (\nabla^2 m(\omega_{i+1})[U_{i+1}, U_{i+1}] + (\partial_1 m_{i+1}) S_{i+1} + (\partial_2 m_{i+1}) S^{\mu}_i)
\end{aligned} \tag{47}$$

Thus if we set

$$\begin{aligned}
\bar{\mu}_{i+1} &= m(\bar{\psi}_{i+1}) \\
G^{\mu}_{i+1} &= (\partial_1 m_{i+1}) G_{i+1} + (\partial_2 m_{i+1}) G^{\mu}_i \\
E^{\mu}_{i+1} &= (\partial_1 m_{i+1}) E_{i+1} + (\partial_2 m_{i+1}) E^{\mu}_i \\
S^{\mu}_{i+1} &= \nabla^2 m(\omega_{i+1})[U_{i+1}, U_{i+1}] + (\partial_1 m_{i+1}) S_{i+1} + (\partial_2 m_{i+1}) S^{\mu}_i,
\end{aligned}$$

then $m\bar{u}_{i+1}$ has the desired independence property and we have recusrive definitions for G^{μ}_{i+1}, E^{μ}_{i+1}, and S^{μ}_{i+1}. This completes the proof of Claim 2.

Before moving on, we remark that it can be easily seen via induction that

$$\begin{aligned}
G^{\mu}_i &= \sum_{j=1}^{i} (\partial_2 m_i) \cdots (\partial_2 m_{j+1})(\partial_1 m_j) G_j \\
&= \sum_{j=1}^{i} (\partial_2 m_i) \cdots (\partial_2 m_{j+1})(\partial_1 m_j) \sum_{l=0}^{j-1} g_l \\
&= \sum_{l=0}^{i-1} \underbrace{\left(\sum_{j=l+1}^{i} (\partial_2 m_i) \cdots (\partial_2 m_{j+1})(\partial_1 m_j) \right)}_{M_{il}} g_l.
\end{aligned} \tag{48}$$

We also have that

$$\|M_{il}\| \le \sum_{j=l+1}^{i} \|\partial_2 m_i\| \cdots \|\partial_2 m_{j+1}\| \|\partial_1 m_j\| \le \sum_{j=l+1}^{i} \delta^{i-j} B \le B \sum_{j=0}^{\infty} \delta^j = \frac{B}{1-\delta}. \tag{49}$$

This uniform constant bound on $\|M_{il}\|$ will be useful later.

We will now bound $\|G^{\mu}_i\|$, $\|E^{\mu}_i\|$, and $\|S^{\mu}_i\|$. When $i = 0$, all of these quantities are 0, which accounts for all of our base cases. We proceed inductively for each one. We first claim $\|G^{\mu}_i\| \le B\mathbf{G} \cdot \frac{1-\delta^i}{1-\delta}$ (recall that $\mathbf{G}$ was chosen so that $\|G_i\| \le \mathbf{G} \le cH\sigma(\sqrt{d}+\sqrt{\log\frac{T}{\gamma}})$:

$$\begin{aligned}\|G^{\mu}_{i+1}\| &\le \|\partial_1 m_{i+1}\|\|G_{i+1}\| + \|\partial_2 m_{i+1}\|\|G^{\mu}_i\| \\ &\le B\mathbf{G} + \delta\cdot B\mathbf{G}\frac{1-\delta^i}{1-\delta} \\ &= B\mathbf{G}\frac{1-\delta^{i+1}}{1-\delta}.\end{aligned}$$

This completes the induction. Since $\delta < 1$, we have $\|G^{\mu}_i\| \le \frac{B\mathbf{G}}{1-\delta} = \widetilde{\mathcal{O}}(\sigma)$ for all i. By the exact same logic and the fact that $\|E_i\| \le H\mathbf{E}_\nabla$ for all i, we also find that $\|E^{\mu}_i\| = \mathcal{O}(\frac{BH\mathbf{E}_\nabla}{1-\delta})$ for all i.

Lastly, define $\mathbf{S} = c'H^2((\sqrt{d}+\sqrt{\log\frac{T}{\gamma}})\sigma + \mathbf{E}_\nabla)$ for some universal constant c' so that $\|S_i\| \le \mathbf{S}$ for all i; this can be done by (45). We claim that $\|S^{\mu}_i\| \le B'\mathbf{S}\frac{1-\delta^i}{1-\delta}$ for all i. We have our base case $i=0$, so we induct. Observe that, by definition of U_{i+1}, we have

$$\begin{aligned}\|U_{i+1}\|^2 &\le (\|G_{i+1}\| + \|E_{i+1}\| + \eta\|S_{i+1}\|)^2 + (\|G^{\mu}_{i+1}\| + \|E^{\mu}_{i+1}\| + \eta\|S^{\mu}_i\|)^2 \\ &\le c(\mathbf{G}^2 + \mathbf{E}^2_\nabla + \eta^2\mathbf{S}^2 + \eta^2\|S^{\mu}_i\|^2) \qquad (50) \\ &= o(\mathbf{S}) \qquad (51)\end{aligned}$$

Here (50) holds by the bounds on $\|G_i\|$, $\|E_i\|$, etc., and the elementary inequality $(x+y+z)^2 = \mathcal{O}(x^2+y^2+z^2)$ for any $x, y, z \ge 0$. Inequality (51) holds since $\mathbf{G}, \mathbf{E}_\nabla, \|S^{\mu}_i\| = \mathcal{O}(\mathbf{S})$, and $\mathbf{S} = o(1)$.

Now, by the recursive definition of S^{μ}_{i+1} and the bounds on $\nabla^2 m$, we have:

$$\begin{aligned}\|S^{\mu}_{i+1}\| &\le C\|U_{i+1}\|^2 + B\mathbf{S} + \delta\|S^{\mu}_i\| \\ &= o(\mathbf{S}) + B\mathbf{S} + \delta\cdot B'\mathbf{S}\frac{1-\delta^i}{1-\delta} \qquad (52) \\ &\le B'\mathbf{S} + \delta\cdot B'\mathbf{S}\frac{1-\delta^i}{1-\delta} \\ &= B'\mathbf{S}\frac{1-\delta^{i+1}}{1-\delta}.\end{aligned}$$

B' can be selected properly in (52) since the $o(\mathbf{S})$ term is vanishingly small compared to $\mathbf{S}$. This completes the induction. Summarizing our results, we have

$$\|E^{\mu}_i\| = \mathcal{O}(\frac{BH\mathbf{E}_\nabla}{1-\delta}), \qquad \|G^{\mu}_i\| = \mathcal{O}(H\sigma(\sqrt{d}+\sqrt{\log\frac{T}{\gamma}})), \qquad \|S^{\mu}_i\| = \widetilde{\mathcal{O}}(\sigma) + \mathcal{O}(\mathbf{E}_\nabla). \qquad (53)$$

With this in mind, we can can now write:

$$\begin{aligned}\psi_{t-H+i} - \psi_t &= \begin{bmatrix}\bar\theta_{t-H+i} - \eta G_{t-H+i} - \eta E_{t-H+i} + \eta^2 S_{t-H+i} - (\bar\theta_t - \eta G_t - \eta E_t + \eta^2 S_t) \\ \bar\mu_{t-H+i-1} - \eta G^{\mu}_{t-H+i-1} - \eta E^{\mu}_{t-H+i-1} + \eta^2 S^{\mu}_{t-H+i-1} - (\bar\mu_{t-1} - \eta G^{\mu}_{t-1} - \eta E^{\mu}_{t-1} + \eta^2 S^{\mu}_{t-1})\end{bmatrix} \\ &= \underbrace{\begin{bmatrix}\bar\theta_{t-H+i} - \bar\theta_t \\ \bar\mu_{t-H+i-1} - \bar\mu_{t-1}\end{bmatrix}}_{\Delta\underline{\psi_i}} + \eta\underbrace{\begin{bmatrix}G_t - G_{t-H+i} \\ G^{\mu}_{t-1} - G^{\mu}_{t-H+i-1}\end{bmatrix}}_{\Delta G_i} + \eta\underbrace{\begin{bmatrix}E_t - E_{t-H+i} \\ E^{\mu}_{t-1} - E^{\mu}_{t-H+i-1}\end{bmatrix}}_{\Delta E_i} + \eta^2\underbrace{\begin{bmatrix}S_t - S_{t-H+i} \\ S^{\mu}_{t-1} - S^{\mu}_{t-H+i-1}\end{bmatrix}}_{\Delta S_i}.\end{aligned}$$

Recall that $v_{i:j}$ denotes the matrix whose columns are v_i through v_j for $i \ge j$:

$$v_{i:j} = \begin{bmatrix}v_i & v_{i-1} & \cdots & v_j\end{bmatrix}.$$

Using this notation, define

$$\underline{\Delta\psi} = \underline{\Delta\psi}_{H-1:0}, \qquad \Delta G = \Delta G_{H-1:0}, \qquad \Delta E = \Delta E_{H-1:0}, \qquad \Delta S = \Delta S_{H-1:0}.$$

Recalling the definition of $\overline{\Delta\psi}$, we then have

$$\overline{\Delta\psi} = \underline{\Delta\psi} + \eta\Delta G + \eta\Delta E + \eta^2\Delta S.$$

Using the same Weyl's inequality calculation as in (19), we the have

$$\|\overline{\Delta\psi}^\dagger\| \leq \frac{\|(\underline{\Delta\psi} + \eta\Delta G)^\dagger\|}{1 - \|(\underline{\Delta\psi} + \eta\Delta G)^\dagger\|(\eta\|\Delta E\| + \eta^2\|\Delta S\|)}. \tag{54}$$

By (45), (53), and the definitions of ΔE and ΔS, we have

$$\|\Delta E\| \leq \|\Delta E\|_F = \mathcal{O}(\frac{BH^{3/2}}{1-\delta}\mathbf{E}_\nabla) \qquad \|\Delta S\| \leq \|\Delta S\|_F = \widetilde{\mathcal{O}}(\sigma) + \mathcal{O}(\mathbf{E}_\nabla). \tag{55}$$

Thus we have reduced our problem to showing that $\underline{\Delta\psi} + \eta\Delta G$ has full rank and bounding $\|(\underline{\Delta\psi} + \eta\Delta G)^\dagger\|$ with high probability. Note that since $\underline{\Delta\psi} + \eta\Delta G \in \mathbb{R}^{2d\times H}$ and $H \geq 2d$, it suffices to show that for any vector $v \in \mathbb{R}^{2d}$ with $\|v\| = 1$, we have

$$\|(\underline{\Delta\psi} + \eta\Delta G)^\top v\| \geq \eta\sigma. \tag{56}$$

First, observe that $\eta\Delta G^\top v$ is Gaussian. Furthermore, for any mean 0 Gaussian vector g and deterministic vector w, by Lemma 9 we have

$$\mathbb{P}(\|w + g\| \geq s) \geq \mathbb{P}(\|g\| \geq s). \tag{57}$$

Thus it suffices to lower bound $\|\Delta G^\top v\|$ with high probability. By the homogeneity of the Gaussian, it suffices to prove the result for $\sigma = 1$.

Again to avoid notational clutter, we will changes indices $t - H + i \mapsto i$. Let $v = (v_1, v_2)$ with $v_1, v_2 \in \mathbb{R}^d$. Observe that $((\Delta G)^\top v)_i = (G_t^\top v_1 + {G_{t-1}^\mu}^\top v_2) + G_i^\top v_1 + {G_{i-1}^\mu}^\top v_2$. Define

$$\widetilde{\Delta G^\top v} = \begin{bmatrix} | \\ G_i^\top v_1 + {G_{i-1}^\mu}^\top v_2 \\ | \end{bmatrix}_{i=0}^{H-1} \in \mathbb{R}^H.$$

We will begin by analyzing $\widetilde{\Delta G^\top v}$ and address the constant offset to this later. First, by definition of G_i and G_i^μ, we have:

$$G_i^\top v_1 + {G_{i-1}^\mu}^\top v_2 = \sum_{l=0}^{i-3} g_l^\top(v_1 + M_{il}^\top v_2) + g_{i-2}^\top(v_1 + (\partial_1 m_{i-1})^\top v_2) + g_{i-1}^\top v_1. \tag{58}$$

We now consider two cases. Recall that we have assume that $\sigma_{\min}(\partial_i m) \geq \alpha$ for all i and some constant $\alpha > 0$.

Case 1: $\|v_1\| \geq \frac{\alpha}{4}$ In this case, we write

$$G_i^\top v_1 + {G_{i-1}^\mu}^\top v_2 = \sum_{l=0}^{i-2} g_l^\top w_{il} + g_{i-1}^\top v_1 \tag{59}$$

with $w_{il} = v_1 + M_{il}^\top v_2$. Define $u = \frac{v_1}{\|v_1\|}$, and for each i, l form the unique decomposition $w_{il} = a_{il}u + w_{il}^\perp$ with $u^\top w_{il}^\perp = 0$. Define $\tilde{g}_l = g_l^\top u$, and note that since $\|u\| = 1$ and $g_l \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$, we have $\tilde{g} \sim \mathcal{N}(0, I_H)$ where $\tilde{g}$ is the vector with entries $\tilde{g}_l$. With this notation, (59) becomes

$$G_i^\top v_1 + {G_{i-1}^\mu}^\top v_2 = \sum_{l=0}^{i-2} a_{il}\tilde{g}_l + \underbrace{\sum_{l=0}^{i-2} g_l^\top w_{il}^\perp}_{Z_i} + \|v_1\|\tilde{g}_{i-1} \tag{60}$$

Note that Z_i is Gaussian and independent of $\tilde{g}$, since

$$\mathbb{E}Z_i\tilde{g}_j = \sum_{l=0}^{i-2}\mathbb{E}[({w_{il}^\perp}^\top g_l)(g_j^\top u)] = \sum_{l=0}^{i-2}{w_{il}^\perp}^\top \mathbb{E}g_l g_j^\top u = \sum_{l=0}^{i-2}\mathbb{1}_{\{l=j\}}{w_{il}^\perp}^\top u = 0$$

by definition of $w_{il}^{\perp}$. Now from (60), we can write

$$\widetilde{\Delta G^\top v} = \tilde{A}\tilde{g} + Z,$$

where $Z \perp\!\!\!\perp \tilde{g}$ and $\tilde{A}$ is lower triangular with 0s along the diagonal, $\|v_1\|$ on the first subdiagonal, and the other entries given by a_{il}. That is,

$$\tilde{A}_{il} = \begin{cases} 0 & l \geq i \\ \|v_1\| & l = i-1 \\ a_{il} & l \leq i-2 \end{cases}.$$

Furthermore, by applying (60) to $i = t$, we have

$$G_t^\top v_1 + {G_{t-1}^{\mu}}^\top v_2 = \sum_{l=0}^{t-1} b_l \tilde{g}_l + \underbrace{\sum_{l=0}^{t-2} g_l^\top w_{tl}^{\perp}}_{z}$$

for some coefficients b_l and z Gaussian and independent of $\tilde{g}$. Let $b \in \mathbb{R}^H$ denote the vector with entries b_l, so that

$$G_t^\top v_1 + {G_{t-1}^{\mu}}^\top v_2 = b^\top \tilde{g} + z.$$

It then follows that

$$\begin{aligned} (\Delta G)^\top v &= (G_t^\top v_1 + {G_{t-1}^{\mu}}^\top v_2)\mathbb{1} - \widetilde{\Delta G^\top v} \\ &= \mathbb{1}(b^\top \tilde{g} + z) - \tilde{A}\tilde{g} - Z \\ &= (\mathbb{1}b^\top - \tilde{A})\tilde{g} + Z', \end{aligned}$$

where $Z' = z\mathbb{1} - Z$ is a mean-0 Gaussian vector independent of $\tilde{g}$.

We now claim that there exists a constant $H_1 = \mathcal{O}(1)$ such that for all $H \geq H_1$, $\sigma_{\sqrt{H}}(\mathbb{1}b^\top - \tilde{A}) \geq \frac{\alpha}{10}$. By Theorem 2.1 of (Zhu et al., 2019), for any k, we have $\sigma_k(\mathbb{1}b^\top - \tilde{A}) \geq \sigma_{k+1}(\tilde{A})$, so it suffices to show that $\sigma_{\sqrt{H}+1}(\tilde{A}) \geq \frac{\alpha}{10}$. By (Horn and Johnson, 2012), Corollary 7.3.6, if M is any matrix and M' is a matrix obtained by deleting a row or column of M, for any k we have $\sigma_k(M) \geq \sigma_k(M')$. Thus if we define A to be the submatrix of $\tilde{A}$ obtained by deleting the first row and last column of $\tilde{A}$, it suffices to show that $\sigma_{\sqrt{H}+1}(A) \geq \frac{\alpha}{10}$.

Observe that A is lower triangular with diagonal entries $\|v_1\| \geq \frac{\alpha}{4}$. Furthermore, all of the entries of A are bounded by a constant: $|A_{ii}| = \|v_1\| \leq 1$, and

$$|a_{il}| \leq \|w_{il}\| \leq \|v_1\| + \|M_{il}\|\|v_2\| \leq 1 + \frac{B}{1-\delta} \leq \frac{2B}{1-\delta} \equiv c,$$

where the penultimate inequality holds by (49) and the final inequality holds by assuming WLOG that $B \geq 1$. It follows that $\|A\|_F \leq c(H-1)$. Furthermore, since A is lower triangular, we have

$$\prod_{i=1}^{H-1} \sigma_i(A) = |\det(A)| = \prod_{i=1}^{H-1} \|v_1\| \geq \left(\frac{\alpha}{4}\right)^{H-1}.$$

Thus we can apply Lemma 11 with $n = H-1$, $\beta = \frac{\alpha}{4}$, $c = \frac{2B}{1-\delta}$, and $k = \sqrt{H}+1$. This implies that there exists a constant $H_1 = \mathcal{O}((\log \frac{B}{\alpha(1-\delta)})^2)$ such that for all $H \geq H_1$, we have $\sigma_{\sqrt{H}+1}(A) \geq \frac{\alpha}{8} \geq \frac{\alpha}{10}$ as desired.

We will now show that $(\Delta G)^\top v$ has a similar decomposition in the other case (when v_1 is small), and after doing so, we can proceed with a unified analysis.

Case 2: $\|v_1\| < \frac{\alpha}{4}$ Observe that since we chose $\|v\| = 1$, we must have

$$\|v_2\| = \sqrt{1 - \|v_1\|^2} \geq \sqrt{1 - (\alpha/4)^2} \geq \sqrt{15}/4$$

since $\alpha \leq 1$. Furthermore, note that

$$\|v_1 + (\partial_1 m_{i-1})^\top v_2\| \geq \sigma_{\min}(\partial_1 m_{i-1})\|v_2\| - \|v_1\| \geq \alpha\frac{\sqrt{15}}{4} - \frac{\alpha}{4} = \frac{\sqrt{15}-1}{4}\alpha.$$

Define $\tilde{v}_{i-2} = v_1 + (\partial_1 m_{i-1})^\top v_2$; the above calculation shows that $\|\tilde{v}_{i-2}\| \geq \frac{\alpha}{2}$. We now proceed similarly to Case 1. As before, we have

$$G_i^\top v_1 + {G_{i-1}^\mu}^\top v_2 = \sum_{l=0}^{i-3} g_l^\top w_{il} + g_{i-2}^\top \tilde{v}_{i-2} + g_{i-1}^\top v_1 \tag{61}$$

with $w_{il} = v_1 + M_{il}^\top v_2$. For each i, l, define $u_i = \tilde{v}_i / \|\tilde{v}_i\|$, and write

$$\begin{aligned} w_{il} &= a_{il} u_l + w_{il}^\perp, \qquad u_l^\top w_{il}^\perp = 0 \\ v_1 &= a_i u_i + v_{1,i}^\perp, \qquad u_i^\top v_{1,i}^\perp = 0. \end{aligned}$$

Set $\tilde{g}_i = g_i^\top u_i$ and let $\tilde{g}$ be the vector with entries $\tilde{g}_i$. Since $\|u_i\| = 1$ and $g_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$, we have $\tilde{g} \sim \mathcal{N}(0, I_H)$. We can now rewrite (61) as

$$G_i^\top v_1 + {G_{i-1}^\mu}^\top v_2 = \sum_{l=0}^{i-3} a_{il}\tilde{g}_l + \|\tilde{v}_{i-2}\|\tilde{g}_{i-2} + a_{i-1}\tilde{g}_{i-1} + \underbrace{\sum_{l=0}^{i-3} g_l^\top w_{il}^\perp + g_{i-1}^\top v_{1,i-1}^\perp}_{Z_i}. \tag{62}$$

Again, Z_i is Gaussian and independent of $\tilde{g}$:

$$\begin{aligned} \mathbb{E} Z_i \tilde{g}_j &= \mathbb{E}[(\sum_{l=0}^{i-3} g_l^\top w_{il}^\perp + g_{i-1}^\top v_{1,i-1}^\perp)(g_j^\top u_j)] \\ &= \sum_{l=0}^{i-3} {w_{il}^\perp}^\top \mathbb{E}[g_l g_j^\top] u_j + v_{1,i-1}^\perp \mathbb{E}[g_{i-1} g_j^\top] u_j \\ &= \sum_{l=0}^{i-3} \mathbb{1}_{l=j} {w_{il}^\perp}^\top u_j + \mathbb{1}_{i-1=j} {v_{1,i-1}^\perp}^\top u_j \qquad (63) \\ &= 0 \end{aligned}$$

where each term in (63) is 0 by definition of $w_{il}^\perp$ and $v_{1,i-1}^\perp$. Furthermore, specializing (62) to $i = t$, we find that there is a vector of constant coefficients b such that

$$G_t^\top v_1 + {G_{t-1}^\mu}^\top v_2 = b^\top \tilde{g} + z$$

where z is a mean-0 Gaussian independent of $\tilde{g}$. Defining $\widetilde{(\Delta G)}^\top v$ as in Case 1, we then have:

$$(\Delta G)^\top v = (\mathbb{1} b^\top - \tilde{A})\tilde{g} + Z'$$

where $Z' = z\mathbb{1} - Z$ is a mean-0 Gaussian vector independent of $\tilde{g}$, and the lower-triangular matrix $\tilde{A}$ is defined by

$$\tilde{A}_{il} = \begin{cases} 0 & l \geq i \\ a_i & l = i-1 \\ \|\tilde{v}_{i-2}\| & l = i-2 \\ a_{il} & l \leq i-3 \end{cases}.$$

We now claim that there is a constant H_2 such that for all $H \geq H_2$, $\sigma_{\sqrt{H}}(\mathbb{1} b^\top - \tilde{A}) \geq \frac{\alpha}{10}$. Define $\hat{A}$ to be $\tilde{A}$ with its first row and lost column deleted. Let $D = \operatorname{diag}(a_i)_{i=1}^{H-1}$ be the diagonal al $\hat{A}$, and finally define A to be the matrix obtained by deleting the first row and last column of $\hat{A} - D$. Note that $A \in \mathbb{R}^{(H-2)\times(H-2)}$ is lower triangular with diagonal entries $\|\tilde{v}_{i-2}\|$, $i = 2, \dots, H-1$. Furthermore, we have:

$$
\begin{aligned}
\sigma_{\sqrt{H}}(\mathbb{1}b^\top - \tilde{A}) &\geq \sigma_{\sqrt{H}+1}(\tilde{A}) && (64)\\
&\geq \sigma_{\sqrt{H}+1}(\hat{A}) && (65)\\
&\geq \sigma_{\sqrt{H}+1}(\hat{A} - D) - \max_i |a_i| && (66)\\
&\geq \sigma_{\sqrt{H}+1}(A) - \max_i |a_i| && (67)\\
&\geq \sigma_{\sqrt{H}+1}(A) - \frac{\alpha}{4}. && (68)
\end{aligned}
$$

Here, (64) follows from (Zhu et al., 2019) Theorem 2.1; (65) and (67) follow (Horn and Johnson, 2012) Corollary 7.3.6; (66) holds by Weyl's inequality for singular values and the fact that $\sigma_1(D) = \max_i |a_i|$; and (68) holds because $|a_i| \leq \|v_1\| \leq \frac{\alpha}{4}$ by definition of a_i. It therefore suffices to show that $\sigma_{\sqrt{H}+1}(A) - \frac{\alpha}{4} \geq \frac{\alpha}{10}$.

Note that A has entries which are bounded by a constant: $|A_{ii}| = \|\tilde{v}_{i-2}\| \leq \|v_1\| + \|\partial_1 m_{i-1}\|\|v_2\| \leq 1 + B$, and

$$
|a_{il}| \leq \|w_{il}\| \leq 1 + \frac{B}{1-\delta} \leq \frac{2B}{1-\delta} \equiv c
$$

as in the previous case, so $\|A\|_F \leq c(H-1)$. Furthermore, since A is lower triangular, its determinant is the product of its diagonal entries, and therefore

$$
\prod_{i=1}^{H-2} \sigma_i(A) = |\det(A)| = \prod_{i=2}^{H-1} \|\tilde{v}_{i-2}\| \geq \left(\frac{\sqrt{15}-1}{4}\alpha\right)^{H-2}.
$$

Thus we can apply Lemma 11 with $n = H-1$, $\beta = \frac{\sqrt{15}-1}{4}\alpha$, and $c = \frac{2B}{1-\delta}$ to conclude that $\sigma_{\sqrt{H}+1}(A) - \frac{\alpha}{4} \geq \frac{\sqrt{15}-1}{8}\alpha - \frac{\alpha}{4} \geq \frac{\alpha}{10}$ for all $H \geq H_2$ with $H_2 = \mathcal{O}((\log \frac{B}{\alpha(1-\delta)})^2)$.

In both cases, we have that $(\Delta G)^\top v = M\tilde{g} + Z'$, where $M \in \mathbb{R}^{H\times H}$ is a matrix with $\sigma_{\sqrt{H}}(M) \geq \frac{\alpha}{10}$, $\tilde{g} \sim \mathcal{N}(0, I_H)$ and Z' is a mean-0 Gaussian with $Z' \perp\!\!\!\perp \tilde{g}$. Thus by Lemma 9 and Lemma 10 with $k = \sqrt{H}$, for all $H \geq \max\{H_1, H_2\}$ we have:

$$
\begin{aligned}
\mathbb{P}(\|(\Delta G)^\top v\| \leq \frac{\alpha}{10}H^{1/4} - r) &= \mathbb{P}(\|M\tilde{g} + Z'\| \leq \alpha H^{1/4} - r)\\
&\leq \mathbb{P}(\|M\tilde{g}\| \leq \alpha H^{1/4} - r)\\
&\leq 2\exp(-c'\frac{r^2}{(\alpha/10)^2}).
\end{aligned}
$$

Then for any $\varepsilon > 0$, setting $r = \frac{\alpha}{10\sqrt{c'}}\sqrt{\log\frac{T}{\gamma} + 2d\log\frac{3}{\varepsilon}}$, we have that

$$
\|(\Delta G)^\top v\| \geq \frac{\alpha}{10}H^{1/4} - \frac{\alpha}{10\sqrt{c'}}\sqrt{\log\frac{T}{\gamma} + 2d\log\frac{3}{\varepsilon}} \qquad \text{with probability } \geq 1 - \frac{2\gamma}{T}\left(\frac{\varepsilon}{3}\right)^{2d} \tag{69}
$$

for any fixed v.

Now let $\{v_i^{\text{net}}\}_{i=1}^{(3/\varepsilon)^{2d}} \subseteq \mathbb{R}^{2d}$ be an ε-net for S^{2d-1}. (An ε-net with $(3/\varepsilon)^{2d}$ elements exists by, e.g., (Vershynin, 2018) Corollary 4.2.13.) By taking a union bound of (69) over each v_i^{net} in the net, we have that (69) holds simultaneously for all v_i^{net} with probability at least $1 - \frac{2\gamma}{T}$. A further union bound shows that this holds over the entire T steps of the trajectory with probability at least $1 - \frac{\gamma}{T}$.

Next, consider any $v \in S^{2d-1}$ and choose v_i^{net} such that $\|v - v_i^{\text{net}}\| \leq \varepsilon$. We have

$$
\|(\Delta G)^\top v\| \geq \|(\Delta G)^\top v_i\| - \|(\Delta G)^\top(v_i - v)\| \geq \frac{\alpha}{10}H^{1/4} - \frac{\alpha}{10\sqrt{c'}}\sqrt{\log\frac{1}{\gamma} + 2d\log\frac{3}{\varepsilon}} - \|\Delta G\|_F\varepsilon. \tag{70}
$$

We can bound $\|\Delta G\|_F$:

$$\|\Delta G\|_F^2 = \sum_{i=0}^{H-1}(\|G_H - G_i\|^2 + \|G^\mu_{H-1} - G^\mu_{i-1}\|^2)$$
$$\leq 2\sum_{i=0}^{H-1}(\|G_H\|^2 + \|G_i\|^2 + \|G^\mu_{H-1}\|^2 + \|G^\mu_{i-1}\|^2).$$

We showed previously that $\|G_i\|, \|G_i^\mu\| = \mathcal{O}(\frac{B}{1-\delta}(\sqrt{d} + \sqrt{\log\frac{T}{\gamma}})H)$ for all i with probability at least $1-\gamma$ over the whole trajectory. Thus

$$\|\Delta G\|_F^2 = \mathcal{O}(\frac{B}{1-\delta}\sqrt{\log\frac{T}{\gamma}}H^{3/2})$$

with probability at least $1-\gamma$. By another union bound, combining this inequality with (70) implies that with probability at least $1-3\gamma$ we have

$$\|(\Delta G)^\top v\| \geq \frac{\alpha}{10}H^{1/4} - \frac{\alpha}{10\sqrt{c'}}\sqrt{\log\frac{1}{\gamma'} + 2d\log\frac{3}{\varepsilon}} - \frac{cB}{1-\delta}\sqrt{\log\frac{T}{\gamma}}H^{3/2}\varepsilon. \tag{71}$$

Setting $\varepsilon = H^{-11/8}$, (71) becomes

$$\|(\Delta G)^\top v\| \geq \frac{\alpha}{10}H^{1/4} - \frac{\alpha}{10\sqrt{c'}}\sqrt{\log\frac{1}{\gamma'} + 2d(\log 3 + \frac{11}{8}\log H)} - \frac{cB}{1-\delta}\sqrt{\log\frac{T}{\gamma}}H^{1/8} \tag{72}$$

$$\geq \frac{\alpha}{10}H^{1/4} - \frac{c_1 B}{1-\delta}\sqrt{d\log\frac{T}{\gamma}}H^{1/8} \tag{73}$$

for some universal constant c_1. The inequality (73) ≥ 1 is quadratic in $H^{1/8}$, and with a simple application of the quadratic formula we see that (73) ≥ 1 whenever $H \geq H_3$ for some $H_3 = \mathcal{O}(\frac{B^8 d^4}{\alpha^8(1-\delta)^8}(\log\frac{T}{\gamma})^4)$.

Combining (73) with (57), we have shown that for $H \geq \max\{2d, H_1, H_2, H_3\} = \mathcal{O}(\frac{B^8}{\alpha^8(1-\delta)^8}(\log\frac{T}{\gamma})^4)$, with probability at least $1-3\gamma$, equation (56) holds at every step in the optimization trajectory. (Recall that it was sufficient to prove this for the special case $\sigma = 1$ by homogeneity.)

We are almost done. Plugging this upper bound into (54), we have

$$\|\overline{\Delta\psi}^\dagger\| \leq \frac{\|(\underline{\Delta\psi} + \eta\Delta G)^\dagger\|}{1 - \|(\underline{\Delta\psi} + \eta\Delta G)^\dagger\|(\eta\|\Delta E\| + \eta^2\|\Delta S\|)}$$
$$\leq \frac{\frac{1}{\eta\sigma}}{1 - \frac{1}{\eta\sigma}(c\eta\frac{BH^{3/2}}{1-\delta}\mathbf{E}_\nabla + \eta^2(\widetilde{\mathcal{O}}(\sigma) + \mathcal{O}(\mathbf{E}_\nabla)))}$$
$$\leq \frac{\frac{1}{\eta\sigma}}{1 - c'\frac{BH^{3/2}}{1-\delta}\frac{\mathbf{E}_\nabla}{\sigma}} \tag{74}$$
$$= \mathcal{O}(\frac{1}{\eta\sigma}), \tag{75}$$

where (74) holds because $\eta(\widetilde{\mathcal{O}}(\sigma) + \mathcal{O}(\mathbf{E}_\nabla)) = o(\mathbf{E}_\nabla)$ and (75) holds provided that $\sigma \geq 2c'\frac{BH^{3/2}\mathbf{E}_\nabla}{1-\delta}$. Given our eventual choices of η and σ and the resulting bound on $\mathbf{E}_\nabla$ from Lemma 4, both of these conditions will indeed hold. Thus we have that $\overline{\Delta\psi}$ is full rank and $\|\overline{\Delta\psi}^\dagger\| = \mathcal{O}(\frac{1}{\eta\sigma})$, as desired. □

Lemma 3. *The long-term Jacobian estimate* $\widehat{\frac{d\mu^*}{d\theta}}$ *from Eq. 2 satisfies*

$$\left\|\widehat{\frac{d\mu^*}{d\theta}} - \frac{d\mu^*}{d\theta}\right\| = \widetilde{\mathcal{O}}(\eta + \mathbf{E}_m) \equiv \mathbf{E}_{\mu^*},$$

where $\mathbf{E}_m$ is the upper bound on the error in estimating the Jacobian of m from Lemma 2.

Proof. Throughout this proof, $m^{(k)}$ denotes $m^{(k)}(\theta_t, \mu_{t-1})$ and μ^* denotes $\mu^*(\theta_t)$. We also seek to evaluate $\frac{dm^{(k)}}{d\theta}$ at the point (θ_t, μ_{t-1}) as well as $\frac{d\mu^*}{d\theta}$ at θ_t. To avoid notational clutter, we will drop the dependence on the time t, so θ denotes θ_t and μ denotes μ_{t-1}.

As $\widehat{\frac{d\mu^*}{d\theta}} = \lim_{k\to\infty} \widehat{\frac{dm^{(k)}}{d\theta}}$ and $\|\frac{dm^{(k)}}{d\theta} - \frac{d\mu^*}{d\theta}\| \to 0$ by Lemma 16, we have

$$\begin{aligned}\|\widehat{\frac{d\mu^*}{d\theta}} - \frac{d\mu^*}{d\theta}\| &= \lim_{k\to\infty} \|\widehat{\frac{dm^{(k)}}{d\theta}} - \frac{d\mu^*}{d\theta}\| \\ &\le \lim_{k\to\infty} \|\widehat{\frac{dm^{(k)}}{d\theta}} - \frac{dm^{(k)}}{d\theta}\| + \|\frac{dm^{(k)}}{d\theta} - \frac{d\mu^*}{d\theta}\| \\ &= \lim_{k\to\infty} \|\widehat{\frac{dm^{(k)}}{d\theta}} - \frac{dm^{(k)}}{d\theta}\|.\end{aligned}$$

Thus it suffices to bound $\|\widehat{\frac{dm^{(k)}}{d\theta}} - \frac{dm^{(k)}}{d\theta}\|$ and take the limit as $k \to \infty$.

Define $B' = \frac{B}{1-\delta}$ and $c_1 = C(1+B')D$. Take c_2 to be a constant such that $\|\mu_t - \mu^*(\theta_t)\| \le c_2 BG(\log\frac{1}{\eta})^2\eta$, which exists by Lemma 18. Finally, define $\beta = c_2 C(1+B')BG(\log\frac{1}{\eta})^2\eta + 2B'\mathbf{E}_m$, and set $\delta' = \delta + \mathbf{E}_m$.

We claim that $\|\frac{dm^{(k)}}{d\theta} - \widehat{\frac{dm^{(k)}}{d\theta}}\| \le c_1 \sum_{i=0}^{k-1} \delta^i(\delta')^{k-1-i} + \beta\frac{1-(\delta')^k}{1-\delta'}$. For $k = 0$, both $\frac{dm^{(k)}}{d\theta}$ and $\widehat{\frac{dm^{(k)}}{d\theta}}$ are 0, so the claim holds trivially. We induct:

$$\begin{aligned}\|\frac{dm^{(k+1)}}{d\theta} - \widehat{\frac{dm^{(k+1)}}{d\theta}}\| &\le \|\partial_1 m(\theta, m^{(k)}) - \partial_1\widehat{m(\theta,\mu)}\| + \|\partial_2 m(\theta, m^{(k)})\frac{dm^{(k)}}{d\theta} - \partial_2\widehat{m(\theta,\mu)}\widehat{\frac{dm^{(k)}}{d\theta}}\| \\ &\le \|\partial_1 m(\theta, m^{(k)}) - \partial_1 m(\theta,\mu)\| + \|\partial_1 m(\theta,\mu) - \partial_1\widehat{m(\theta,\mu)}\| \\ &+ \|\partial_2 m(\theta, m^{(k)}) - \partial_2 m(\theta,\mu)\|\|\frac{dm^{(k)}}{d\theta}\| + \|\partial_2 m(\theta,\mu) - \partial_2\widehat{m(\theta,\mu)}\|\|\frac{dm^{(k)}}{d\theta}\| \\ &+ \|\partial_2\widehat{m(\theta,\mu)}\|\|\frac{dm^{(k)}}{d\theta} - \widehat{\frac{dm^{(k)}}{d\theta}}\| \\ &\le C\|m^{(k)} - \mu\| + \mathbf{E}_m + C\|m^{(k)} - \mu\|B' + \mathbf{E}_m B' + (\delta + \mathbf{E}_m)\|\frac{dm^{(k)}}{d\theta} - \widehat{\frac{dm^{(k)}}{d\theta}}\| \qquad (76) \\ &= C(1+B')\|m^{(k)} - \mu\| + (1+B')\mathbf{E}_m + (\delta + \mathbf{E}_m)\|\frac{dm^{(k)}}{d\theta} - \widehat{\frac{dm^{(k)}}{d\theta}}\| \\ &\le C(1+B')(\|m^{(k)} - \mu^*\| + \|\mu^* - \mu\|) + (1+B')\mathbf{E}_m + (\delta + \mathbf{E}_m)\|\frac{dm^{(k)}}{d\theta} - \widehat{\frac{dm^{(k)}}{d\theta}}\| \\ &\le C(1+B')(D\delta^k + c_2 BG(\log\frac{1}{\eta})^2\eta) + 2B'\mathbf{E}_m + \delta'\|\frac{dm^{(k)}}{d\theta} - \widehat{\frac{dm^{(k)}}{d\theta}}\| \qquad (77) \\ &\le c_1\delta^k + \beta + \delta'(c_1\sum_{i=0}^{k-1}\delta^i(\delta')^{k-1-i} + \beta\frac{1-(\delta')^k}{1-\delta'}) \qquad (78) \\ &= c_1(\delta^k + \sum_{i=0}^{k-1}\delta^i(\delta')^{k-i}) + \beta(1 + \frac{\delta' - (\delta')^{k+1}}{1-\delta'}) \\ &= c_1\sum_{i=0}^{k}\delta^i(\delta')^{k-i} + \beta\frac{1-(\delta')^{k+1}}{1-\delta'}.\end{aligned}$$

In the above, (76) holds by Assumption 5; (77) holds by Lemma 14; and 78 holds by definition of c_1, c_2, β, δ', and the inductive hypothesis. Since $\mathbf{E}_m \geq 0$, we have that $\delta' \geq \delta$. Furthermore, since $\mathbf{E}_m = o(1)$, we may assume that $0 < \delta' < 1$. It follows that

$$\begin{aligned}\|\frac{dm^{(k)}}{d\theta} - \widehat{\frac{dm^{(k)}}{d\theta}}\| &\leq c_1 \sum_{i=0}^{k-1} \delta^i (\delta')^{k-1-i} + \beta \frac{1-(\delta')^k}{1-\delta'} \\ &\leq c_1 k (\delta')^{k-1} + (c_2 C(1+B')BG(\log\frac{1}{\eta})^2\eta + (1+B')\mathbf{E}_m)\frac{1}{1-\delta'} \\ &= \mathcal{O}\left(k(\delta')^k + \frac{CB^2G}{(1-\delta)^2}(\log\frac{1}{\eta})^2\eta + \frac{B}{(1-\delta)^2}\mathbf{E}_m\right).\end{aligned}$$

Taking $k \to \infty$, we find that

$$\|\widehat{\frac{d\mu^*}{d\theta}} - \frac{d\mu^*}{d\theta}\| = \mathcal{O}\left(\frac{CB^2G}{(1-\delta)^2}(\log\frac{1}{\eta})^2\eta + \frac{B}{(1-\delta)^2}\mathbf{E}_m\right) \equiv \mathbf{E}_{\mu^*}$$

as desired. Substituting the expression for $\mathbf{E}_m$, we see that the $(\log\frac{1}{\eta})^2\eta$ term does not contribute to leading order, so we have

$$\mathbf{E}_{\mu^*} = \mathcal{O}\left(\frac{C\sqrt{H}B^3G^2}{(1-\delta)^2}(\log\frac{1}{\eta})^4\frac{\eta}{\sigma} + \frac{B^3GH}{(1-\delta)^2}(\log\frac{1}{\eta})^2\frac{\varepsilon}{\eta\sigma^2}\right). \tag{79}$$

□

Lemma 4. *The estimator* $\widehat{\nabla\mathcal{L}_t^*}$ *from Eq.* (1) *satisfies*

$$\|\widehat{\nabla\mathcal{L}_t^*} - \nabla\mathcal{L}_t^*\| = \widetilde{\mathcal{O}}(\eta + \mathbf{E}_{\mu^*}),$$

where $\mathbf{E}_{\mu^*}$ *is the error bound on* $\widehat{\frac{d\mu^*}{d\theta}}$ *from Lemma 3.*

Proof. We have

$$\nabla\mathcal{L}^*(\theta) = \overbrace{\int \nabla_\theta \ell(z,\theta) p(z,\mu^*(\theta))\,dz}^{\nabla_1\mathcal{L}^*(\theta)} + \overbrace{\int \ell(z,\theta)\frac{d\mu^*}{d\theta}^\top \nabla_\mu p(z,\mu^*(\theta))\,dz}^{\nabla_2\mathcal{L}^*(\theta)}$$

$$\hat{\nabla}\mathcal{L}^*(\theta) = \underbrace{\int \nabla_\theta \ell(z,\theta) p(z,\hat{\mu})\,dz}_{\hat{\nabla}_1\mathcal{L}^*(\theta)} + \underbrace{\int \ell(z,\theta)\widehat{\frac{d\mu^*}{d\theta}}^\top \nabla_\mu p(z,\hat{\mu})\,dz}_{\hat{\nabla}_2\mathcal{L}^*(\theta)}$$

We write $\|\nabla\mathcal{L}^* - \hat{\nabla}\mathcal{L}^*\| \leq \|\nabla_1\mathcal{L}^* - \hat{\nabla}_1\mathcal{L}^*\| + \|\nabla_2\mathcal{L}^* - \hat{\nabla}_2\mathcal{L}^*\|$ and bound each of these terms separately. For the remainder of this proof, we will assume that $\|\hat{\mu} - \mu^*\| = o(1)$. By the result of Lemma 18, this will be the case when $\eta = o(1)$.

Let $\sigma_0^2 = \|\Sigma\|$, and let $B^d(R)$ denote the Euclidean ball in $\mathbb{R}^d$ of radius R. For the first term, we have

$$\begin{aligned}
\|\nabla_1 \mathcal{L}^* - \hat{\nabla}_1 \mathcal{L}^*\| &\le \ell_{\max} \int |p(z,\mu^*) - p(z,\hat{\mu})|\, dz \\
&\le \ell_{\max} \left[\int_{\|z-\hat{\mu}\|\le R} |p(z,\mu^*) - p(z,\hat{\mu})|\, dz + \int_{\|z-\hat{\mu}\|>R} p(z,\hat{\mu})\, dz + \int_{\|z-\hat{\mu}\|>R} p(z,\mu^*)\, dz \right] \\
&\le \ell_{\max} \left[L_{p,2}\|\mu^* - \hat{\mu}\|\mathrm{vol}(B^d(R)) + \int_{\|z-\hat{\mu}\|>R} p(z,\hat{\mu})\, dz + \int_{\|z-\mu^*\|>R-\|\hat{\mu}-\mu^*\|} p(z,\mu^*)\, dz \right] \quad (80) \\
&= \ell_{\max}[L_{p,2}\|\mu^* - \hat{\mu}\|\mathrm{vol}(B^d(R)) \\
&\quad + \mathbb{P}_{z\sim\mathcal{N}(\hat{\mu},\Sigma)}(\|z-\hat{\mu}\| \ge R) + \mathbb{P}_{z\sim\mathcal{N}(\mu^*,\Sigma)}(\|z-\mu^*\| \ge R - \|\hat{\mu}-\mu^*\|)] \\
&\le \ell_{\max}[6L_{p,2}R^d\|\mu^* - \hat{\mu}\| \\
&\quad + c_1 \exp\{-c_2(R - \sigma_0\sqrt{d})^2/\sigma_0^2\}] + c_1 \exp\{-c_2(R - \sigma_0\sqrt{d} - \|\hat{\mu}-\mu^*\|)^2/\sigma_0^2\} \quad (81) \\
&= \mathcal{O}(R^d\|\mu^* - \hat{\mu}\| + \exp\{-c_2(R - \sigma_0\sqrt{d} - \|\hat{\mu}-\mu^*\|)^2/\sigma_0^2\}) \quad (82)
\end{aligned}$$

for any $R \ge \sigma_0\sqrt{d} + \|\hat{\mu} - \mu^*\|$. In the above, (80) holds because

$$\|z - \hat{\mu}\| > R \implies \|z - \mu^*\| > \|z - \hat{\mu}\| - \|\hat{\mu} - \mu^*\| > R - \|\hat{\mu} - \mu^*\|.$$

Equation (81) holds by the inequality $\mathrm{vol}(B^d(R)) \le 6R^d$, and by Lemma 7. If we then set

$$R = \sigma_0\sqrt{d} + \|\hat{\mu} - \mu^*\| + \frac{\sigma_0}{\sqrt{c_2}}\sqrt{\log\frac{1}{\|\hat{\mu} - \mu^*\|}},$$

substituting into (82) yields

$$\begin{aligned}
\|\nabla_1 \mathcal{L}^* - \hat{\nabla}_1 \mathcal{L}^*\| &= \mathcal{O}(3^d[(\sigma_0\sqrt{d})^d + \|\hat{\mu}-\mu^*\|^d + (\frac{\sigma_0}{\sqrt{c_2}}(\log\frac{1}{\|\hat{\mu}-\mu^*\|})^{1/2})^d]\|\hat{\mu}-\mu^*\| + \|\hat{\mu}-\mu^*\|) \quad (83) \\
&= \mathcal{O}\left(\left(\log\frac{1}{\|\hat{\mu}-\mu^*\|}\right)^{d/2} \|\hat{\mu}-\mu^*\| \right).
\end{aligned}$$

Equation (83) holds by the elementary inequality $(a+b+c)^d \le 3^d(a^d + b^d + c^d)$ for any $a, b, c, d \ge 0$.

The bound on the second gradient term is similar to the first. First, for the Gaussian density p, note that $\nabla_\mu p(z,\mu) = \Sigma^{-1}(\mu - z)p(z,\mu)$. Using this fact, we have

$$\begin{aligned}
\|\nabla_2 \mathcal{L}^* - \hat{\nabla}_2 \mathcal{L}^*\| &\le \ell_{\max} \int \|\frac{d\mu^*}{d\theta}^\top \nabla_\mu p(z,\mu^*) - \widehat{\frac{d\mu^*}{d\theta}}^\top \nabla_\mu p(z,\hat{\mu})\|\, dz \\
&\le \ell_{\max} \left[\|\frac{d\mu^*}{d\theta}^\top\| \int \|\nabla_\mu p(z,\mu^*) - \nabla_\mu p(z,\hat{\mu})\|\, dz + \|\frac{d\mu^*}{d\theta}^\top - \widehat{\frac{d\mu^*}{d\theta}}^\top\| \int \|\nabla_\mu p(z,\hat{\mu})\|\, dz \right] \\
&\le \ell_{\max} \left[\frac{B}{1-\delta} L_{\nabla_\mu p,2}\|\mu^* - \hat{\mu}\|\mathrm{vol}(B^d(R)) + \int_{\|z-\hat{\mu}\|>R} \|\nabla_\mu p(z,\hat{\mu})\|\, dz \right. \\
&\quad \left. + \int_{\|z-\mu^*\|>R-\|\mu^*-\hat{\mu}\|} \|\nabla_\mu p(z,\mu^*)\|\, dz + \|\frac{d\mu^*}{d\theta} - \widehat{\frac{d\mu^*}{d\theta}}\| \int \|\nabla_\mu p(z,\hat{\mu})\|\, dz \right] \\
&= \mathcal{O}\left(R^d\|\mu^* - \hat{\mu}\| + \int_{\|z-\mu^*\|>R-\|\mu^*-\hat{\mu}\|} \|\nabla_\mu p(z,\mu^*)\|\, dz + \|\frac{d\mu^*}{d\theta} - \widehat{\frac{d\mu^*}{d\theta}}\| \|\Sigma^{-1/2}\|\sqrt{d} \right). \quad (84)
\end{aligned}$$

Equation (84) follows by applying Lemma 6 to $\int \|\nabla_\mu p\|$. We bound the integral in the last line separately. For any $r > \sigma_0\sqrt{d}$, we have

$$\begin{aligned}
\int_{\|z-\mu^*\|>r} \|\nabla_\mu p(z,\mu^*)\|\, dz &= \int_{\|z-\mu^*\|>r} \|\Sigma^{-1}(\mu^* - z)\| p(z,\mu^*)\, dz \\
&= \mathbb{E}_{z\sim\mathcal{N}(\mu^*,\Sigma)}[\|\Sigma^{-1}(z-\mu^*)\| \cdot \mathbb{1}\{\|z-\mu^*\| > r\}] \\
&\leq \sqrt{\mathbb{E}_{z\sim\mathcal{N}(\mu^*,\Sigma)}[\|\Sigma^{-1}(z-\mu^*)\|^2] \cdot \mathbb{E}_{z\sim\mathcal{N}(\mu^*,\Sigma)}[\mathbb{1}\{\|z-\mu^*\| > r\}^2]} \qquad (85) \\
&\leq \sqrt{\|\Sigma^{-1/2}\|^2 \mathbb{E}_{z\sim\mathcal{N}(\mu^*,\Sigma)}[\|\Sigma^{-1/2}(z-\mu^*)\|^2] \cdot \mathbb{P}_{z\sim\mathcal{N}(\mu^*,\Sigma)}(\|z-\mu^*\| > r)} \\
&= \|\Sigma^{-1/2}\| \sqrt{\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\|z\|^2] \cdot c_1 \exp\{-c_2(r-\sigma_0\sqrt{d})^2/\sigma_0^2\}} \qquad (86) \\
&= \|\Sigma^{-1/2}\| \sqrt{d} \cdot \sqrt{c_1} \exp\{-c_2(r-\sigma_0\sqrt{d})^2/2\sigma_0^2\}. \qquad (87)
\end{aligned}$$

Inequality (85) holds by the Cauchy-Schwarz inequality. Equation (86) holds because

$$z \sim \mathcal{N}(\mu^*, \Sigma) \implies \Sigma^{-1}(z-\mu^*) \sim \mathcal{N}(0, I_d),$$

and by Lemma 7. Finally, equation (87) holds because $\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}\|z\|^2 = d$. We can now plug (87) into (84):

$$\|\nabla_2\mathcal{L}^* - \hat{\nabla}_2\mathcal{L}^*\| = \mathcal{O}\left(R^d\|\mu^* - \hat{\mu}\| + \exp\{-c_2(R - \|\mu^* - \hat{\mu}\| - \sigma_0\sqrt{d})^2/2\sigma_0^2\} + \|\frac{d\mu^*}{d\theta} - \widehat{\frac{d\mu^*}{d\theta}}\|\|\Sigma^{-1/2}\|\sqrt{d}\right).$$

If we take $R = \sigma_0\sqrt{d} + \|\hat{\mu} - \mu^*\| + \frac{\sqrt{2}\sigma_0}{\sqrt{c_2}}\sqrt{\log\frac{1}{\|\hat{\mu}-\mu^*\|}}$, then by the same logic as was used in Equation (83), we obtain

$$\|\nabla_2\mathcal{L}^* - \hat{\nabla}_2\mathcal{L}^*\| = \mathcal{O}\left(\left(\log\frac{1}{\|\hat{\mu}-\mu^*\|}\right)^{d/2}\|\hat{\mu}-\mu^*\| + \|\frac{d\mu^*}{d\theta} - \widehat{\frac{d\mu^*}{d\theta}}\|\|\Sigma^{-1/2}\sqrt{d}\right).$$

Thus the bound on $\|\nabla_1\mathcal{L}^* - \hat{\nabla}_1\mathcal{L}^*\|$ can be absorbed into the bound on $\|\nabla_2\mathcal{L}^* - \hat{\nabla}_2\mathcal{L}^*\|$, and we have

$$\|\nabla\mathcal{L}^* - \hat{\nabla}\mathcal{L}^*\| = \mathcal{O}\left(\left(\log\frac{1}{\|\hat{\mu}-\mu^*\|}\right)^{d/2}\|\hat{\mu}-\mu^*\| + \|\frac{d\mu^*}{d\theta} - \widehat{\frac{d\mu^*}{d\theta}}\|\|\Sigma^{-1/2}\|\sqrt{d}\right). \qquad (88)$$

Now, when we take $\theta = \theta_t$ (so $\mu^* = \mu^*(\theta_t)$ and we are evaluating $\frac{d\mu^*}{d\theta}$ at θ_t) and $\hat{\mu} = \hat{\mu}_t$, by Lemma 18, for $t \geq \log_\delta \eta$, we have

$$\|\hat{\mu}_t - \mu^*(\theta_t)\| \leq \|\hat{\mu}_t - \mu_t\| + \|\mu_t - \mu^*(\theta_t)\| \leq \mathcal{O}(\boldsymbol{\varepsilon} + (\log\frac{1}{\eta})^2\eta).$$

Substituting this into (88) and using the definition of $\mathbf{E}_{\mu^*}$, we have

$$\begin{aligned}
\|\nabla\mathcal{L}^* - \hat{\nabla}\mathcal{L}^*\| &= \mathcal{O}\left(\left(\log\frac{1}{\boldsymbol{\varepsilon} + \left(\log\frac{1}{\eta}\right)^2\eta}\right)^{d/2}\left(\boldsymbol{\varepsilon} + \left(\log\frac{1}{\eta}\right)^2\eta\right) + \sqrt{d}\|\Sigma^{-1/2}\|\mathbf{E}_{\mu^*}\right) \\
&= \widetilde{\mathcal{O}}(\boldsymbol{\varepsilon} + \eta) + \mathcal{O}(\sqrt{d}\|\Sigma^{-1/2}\|\mathbf{E}_{\mu^*}).
\end{aligned}$$

Since $\boldsymbol{\varepsilon} = \mathcal{O}(\eta)$, we obtain the desired result. From the expression (79) for $\mathbf{E}_{\mu^*}$, we see that the $\widetilde{\mathcal{O}}(\eta)$ term does not contribute to leading order, and we obtain

$$\|\nabla\mathcal{L}^* - \hat{\nabla}\mathcal{L}^*\| = \mathcal{O}\left(\frac{CB^3G^2\sqrt{Hd}\|\Sigma^{-1/2}\|}{(1-\delta)^2}(\log\frac{1}{\eta})^4\frac{\eta}{\sigma} + \frac{B^3GH\sqrt{d}\|\Sigma^{-1/2}\|}{(1-\delta)^2}(\log\frac{1}{\eta})^2\frac{\boldsymbol{\varepsilon}}{\eta\sigma^2}\right) \equiv \mathbf{E}_\nabla. \qquad (89)$$

Note that this matches the definition of $\mathbf{E}_\nabla$ given in (40). □

Lemma 5. *Let h be any L-smooth function and let $\widehat{\nabla h}$ be a gradient oracle with bounded error: $\|\widehat{\nabla h}(x) - \nabla h(x)\| \leq \mathbf{e}$, and assume that $\mathbf{e} = o(1)$. Then for η sufficiently small, the iterates x_t of gradient descent with gradient oracle $\widehat{\nabla h}$ satisfy*

$$\min_{1\leq t\leq T} \|\nabla h(x_t)\|^2 = \mathcal{O}\left(\frac{1}{T\eta} + \mathbf{e}\right).$$

Proof. We require the additional assumptions that $\|\nabla h(x)\| \leq G$ and that $|h(x)| \leq h_{\max}$ for all x, and that $G, h_{\max} = \mathcal{O}(1)$.

Since h is L-smooth, we have $h(y) \leq h(x) + \nabla h(x)^\top (y-x) + \frac{L}{2}\|x-y\|^2$ for any x, y. Define $e_t = \widehat{\nabla} h(x_t) - \nabla h(x_t)$, so $\widehat{\nabla h}(x_t) = \nabla h(x_t) + e_t$. Taking $x = x_t$ and $y = x_{t+1}$, we have

$$\begin{aligned}
h(x_{t+1}) &\leq h(x_t) + \nabla h(x_t)^\top (x_t - \eta(\nabla h(x_t) + e_t) - x_t) + \frac{L}{2}\|x_t - \eta(\nabla h(x_t) + e_t) - x_t\|^2 \\
&\leq h(x_t) - \eta\|\nabla h(x_t)\|^2 + \eta\|\nabla h(x_t)\|\|e_t\| + \eta^2 L\|\nabla h(x_t)\|^2 + \eta^2 L\|e_t\|^2 \qquad (90)\\
&\leq h(x_t) - \eta\|\nabla h(x_t)\|^2 + \eta G\mathbf{e} + \eta^2 L\|\nabla h(x_t)\|^2 + \eta^2 L\mathbf{e}^2 \\
&= h(x_t) + (\eta^2 L - \eta)\|\nabla h(x_t)\|^2 + \eta G\mathbf{e} + \eta^2 L\mathbf{e}^2.
\end{aligned}$$

Here (90) holds by the Cauchy-Schwarz inequality. Since $\eta = o(1)$, we may assume that $\eta - \eta^2 L > 0$. Rearranging, it follows that

$$\|\nabla h(x_t)\|^2 \leq \frac{h(x_t) - h(x_{t+1}) + \eta G\mathbf{e} + \eta^2 L\mathbf{e}^2}{\eta - \eta^2 L}. \qquad (91)$$

We now sum (91) from $t = 1$ to T. This yields

$$\begin{aligned}
T \min_{1\leq t\leq T} \|\nabla h(x_t)\|^2 &\leq \sum_{t=1}^T \|\nabla h(x_t)\|^2 \\
&\leq \sum_{t=1}^T \frac{h(x_t) - h(x_{t+1}) + \eta G\mathbf{e} + \eta^2 L\mathbf{e}^2}{\eta - \eta^2 L} \\
&= \frac{h(x_1) - h(x_{T+1})}{\eta - \eta^2 L} + T\cdot\frac{\eta G\mathbf{e} + \eta^2 L\mathbf{e}^2}{\eta - \eta^2 L} \\
&\leq \frac{2h_{\max}}{\eta - \eta^2 L} + T\cdot\frac{\eta G\mathbf{e} + \eta^2 L\mathbf{e}^2}{\eta - \eta^2 L} \\
&= \mathcal{O}\left(\frac{h_{\max}}{\eta} + GT\mathbf{e}\right). \qquad (92)
\end{aligned}$$

Here (92) holds since $\eta = o(1)$ and $\mathbf{e} = o(1)$. Dividing both sides of (92) by T yields the desired result. □

Theorem 1. *Let T be the number of deployments of Stateful PerfGD, and for each t let $\nabla\mathcal{L}_t^* = \nabla\mathcal{L}^*(\theta_t)$. Then for any $\gamma > 0$, there exist intervals $[\eta_{\min}, \eta_{\max}]$ and $[\sigma_{\min}, \sigma_{\max}]$ (which depend on T and the estimation error $\boldsymbol{\varepsilon}$) such that for any learning rate η in the former and perturbation size σ in the latter interval, with probability at least $1-\gamma$, the iterates of Stateful PerfGD satisfy*

$$\min_{1\leq t\leq T} \|\nabla\mathcal{L}_t^*\|^2 = \widetilde{\mathcal{O}}(T^{-1/5} + \boldsymbol{\varepsilon}^{1/5}).$$

Proof. First, we remark that in order for Lemma 18 to hold, we need a "warm-up" phase of length $\log\frac{1}{\eta}$. We will always take $\eta = \Omega(T^{-2/5})$, in which case this warm-up phase has length $\mathcal{O}(\log T^{2/5}) = \mathcal{O}(\log T) = o(T)$. This does not change the asymptotic length of the trajectory, so we will simply ignore it in the following calculations. We will also assume that all of the required high-probability events hold from each of the previous lemmas, making the following statements hold with probability at least $1 - \mathcal{O}(\gamma)$.

We split into two (very similar) cases. First, we consider when $\boldsymbol{\varepsilon} \geq \frac{1}{T}$.

Suppose that for $s \le t$, we have that $\|\hat{\nabla}\mathcal{L}^*(\theta_s) - \nabla\mathcal{L}^*(\theta_s)\| \le \mathbf{E}_\nabla$ with

$$\mathbf{E}_\nabla = \mathcal{O}\left(\frac{CB^3G^2\sqrt{Hd}\|\Sigma^{-1/2}\|}{(1-\delta)^2}(\log\frac{1}{\eta})^4\frac{\eta}{\sigma} + \frac{B^3GH\sqrt{d}\|\Sigma^{-1/2}\|}{(1-\delta)^2}(\log\frac{1}{\eta})^2\frac{\varepsilon}{\eta\sigma^2}\right). \tag{93}$$

(Again, note that if $s \le \log\frac{1}{\eta}$, then we replace $\hat{\nabla}\mathcal{L}^*(\theta_s)$ and $\nabla\mathcal{L}^*(\theta_s)$ with 0 and the above bound holds trivially.) Since $\mathbf{E}_\nabla = o(1)$ and $\|\hat{\nabla}\mathcal{L}^*(\theta_s)\| \le G + \mathbf{E}_\nabla$, we have $\|\hat{\nabla}\mathcal{L}^*(\theta_s)\| \le cG$ for some absolute constant c. Furthermore, if we require that

$$\sigma \ge 2c'\frac{BH^{3/2}\mathbf{E}_\nabla}{1-\delta}, \tag{94}$$

then Lemma 2 holds. Then we have the chain Lemma 2 $\Rightarrow$ Lemma 3 $\Rightarrow$ Lemma 4, and in particular Lemma 4 holds with the same $\mathbf{E}_\nabla$ as in (93). Inductively, we see that $\|\hat{\nabla}\mathcal{L}^*(\theta_t) - \nabla\mathcal{L}^*(\theta_t)\| \le \mathbf{E}_\nabla$ for all $1 \le t \le T$.

We can now apply Lemma 5 with $\mathbf{e} = \mathbf{E}_\nabla + \mathcal{O}(\sigma\sqrt{\log\frac{T}{\gamma}})$. (The second term accounts for the fact that the perturbation g_t must be included in $\mathbf{e}$, and $\|g_t\| = \mathcal{O}(\sigma\sqrt{\log\frac{T}{\gamma}})$.) This yields

$$\min_{1\le t\le T}\|\nabla\mathcal{L}_t^*\|^2 = \mathcal{O}\left(\frac{\ell_{\max}}{T\eta} + G(\mathbf{E}_\nabla + \sigma\sqrt{\log\frac{T}{\gamma}})\right) = \mathcal{O}\left(\frac{\ell_{\max}\varepsilon}{\eta} + G\sqrt{\log\frac{T}{\gamma}}\cdot\sigma\right) \tag{95}$$

where the second bound holds by condition (94) and $\varepsilon \ge 1/T$.

Next, let us analyze the condition (94); it takes the form $\sigma \ge c_1(\log\frac{1}{\eta})^4\frac{\eta}{\sigma} + c_2(\log\frac{1}{\eta})^2\frac{\varepsilon}{\eta\sigma^2}$. Dividing both sides of this inequality by σ and setting $\eta = \sqrt{\varepsilon/\sigma}$, (94) holds if

$$c_1(\log\frac{1}{\eta})^4\frac{\eta}{\sigma^2} + c_2(\log\frac{1}{\eta})^2\frac{\varepsilon}{\eta\sigma^3} \le (c_1+c_2)(\log\frac{1}{\eta})^4\frac{\varepsilon^{1/2}}{\sigma^{5/2}} \le (c_1+c_2)(\log\frac{1}{\varepsilon})^4\frac{\varepsilon^{1/2}}{\sigma^{5/2}} \le 1.$$

The rightmost inequality holds when $\sigma \ge (c_1+c_2)^{2/5}(\log\frac{1}{\varepsilon})^{8/5}\varepsilon^{1/5}$. Furthermore, the expressions for the coefficients c_1 and c_2 are given by

$$c_1 = \frac{2c'BH^{3/2}}{1-\delta}\cdot c\frac{CB^3G^2\sqrt{Hd}\|\Sigma^{-1/2}\|}{(1-\delta)^2} = \frac{c''B^4G^2H^2\sqrt{d}\|\Sigma^{-1/2}\|}{(1-\delta)^3}$$

$$c_2 = \frac{2c'BH^{3/2}}{1-\delta}\cdot c\frac{B^3GH\sqrt{d}\|\Sigma^{-1/2}\|}{(1-\delta)^2} = \frac{c''B^4GH^{5/2}\sqrt{d}\|\Sigma^{-1/2}\|}{(1-\delta)^3}.$$

It follows that

$$(c_1+c_2)^{2/5} = \mathcal{O}\left(\frac{B^{8/5}G^{4/5}Hd^{1/5}\|\Sigma^{-1}\|^{1/5}}{(1-\delta)^{6/5}}\right),$$

which finally yields that (94) holds for

$$\sigma = \Omega\left(\frac{B^{8/5}G^{4/5}Hd^{1/5}\|\Sigma^{-1}\|^{1/5}}{(1-\delta)^{6/5}}(\log\frac{1}{\varepsilon})^{8/5}\varepsilon^{1/5}\right). \tag{96}$$

Finally, we set $\sigma = c(\log\frac{1}{\varepsilon})^{8/5}\varepsilon^{1/5}$ and $\eta = \sqrt{\varepsilon/\sigma}$. Observe that in this case, $\frac{\varepsilon}{\eta} = \varepsilon^{1/2}\sigma^{1/2} = \widetilde{\mathcal{O}}(\varepsilon^{3/5}) = o(\sigma)$, so the ε/η term in (95) can be ignored. Recalling the fact that we must choose $H = \Theta(\frac{B^8d^4}{\alpha^8(1-\delta)^8}(\log\frac{T}{\gamma})^4)$ in order for Lemma 19 to hold, we have

$$\begin{aligned}\min_{1\le t\le T}\|\nabla\mathcal{L}_t^*\|^2 &= \mathcal{O}\left(\frac{B^{8/5}G^{9/5}Hd^{1/5}\|\Sigma^{-1}\|^{1/5}}{(1-\delta)^{6/5}}\sqrt{\log\frac{T}{\gamma}}(\log\frac{1}{\varepsilon})^{8/5}\varepsilon^{1/5}\right)\\ &= \mathcal{O}\left(\frac{B^{9.6}G^{1.8}d^{4.2}\|\Sigma^{-1}\|^{0.2}}{\alpha^8(1-\delta)^{9.2}}(\log\frac{T}{\gamma})^{4.5}(\log\frac{1}{\varepsilon})^{1.6}\cdot\varepsilon^{1/5}\right)\\ &= \widetilde{\mathcal{O}}(\varepsilon^{1/5}).\end{aligned}$$

This completes the case when $\varepsilon \geq \frac{1}{T}$. Otherwise, we have $\varepsilon < \frac{1}{T}$. Starting from (93), WLOG we can replace each occurence of ε with $\frac{1}{T}$ and all of the bounds will still hold, so in this case we get

$$\mathcal{O}\left(\frac{B^{9.6}G^{1.8}d^{4.2}\|\Sigma^{-1}\|^{0.2}}{\alpha^8(1-\delta)^{9.2}}(\log\frac{T}{\gamma})^{4.5}(\log T)^{1.6}\cdot T^{-1/5}\right) = \widetilde{\mathcal{O}}(T^{-1/5}).$$

Since we are always in one case or the other, we always have

$$\min_{1\leq t\leq T}\|\nabla\mathcal{L}_t^*\|^2 = \widetilde{\mathcal{O}}(T^{-\frac{1}{5}} + \varepsilon^{\frac{1}{5}}).$$

The fact that there are interval of nonzero width for η and σ follows from the fact that we can multiply or divide both of these by constants close to 1 and not change any of the asymptotics (since the constants can be chosen such that (94) still holds). This completes the proof. □

C EXPERIMENT DETAILS

For all of the experiments, we did a grid search over the relevant parameters for each method, then chose the best results for that method. The parameters we considered were:

RGD → learning rate (lr)
DFO → learning rate (lr), wait, perturbation size (ps)
PerfGD → learning rate (lr), wait, horizon (H)
SPGD → learning rate (lr), perturbation size (ps), horizon (H)

The grid ranges for each parameter were as follows:

- lr $\in \{10^{-k/2} : k = 1, \dots, 6\}$
- wait $\in \{1, 5, 10, 20\}$
- DFO ps $\in \{10^{-k/2} : k = 0, \dots, 3\}$
- SPGD ps $\in \{0\} \cup \{10^{-k/2} : k = 0, \dots, 3\}$
- PGD $H \in \{d,\, d+1,\, \dots, 2d,\, \infty\}$
- SPGD $H \in \{2d,\, 2d+1,\, \dots,\, 3d,\, \infty\}$

For each experiment, we specify the dimension d. We also require that $\theta \in [-R, R]^d$ for some R. If any of the optimization methods took θ outside of this constraint set, we simply clamped θ back to the required range.

Rather than having deterministically bounded error on the mean estimates $\hat{\mu}_t$, we take $\hat{\mu}_t = \mu_t + e_t$, where $e_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\text{err}}^2 I)$ are Gaussian error terms which would arise from taking $\hat{\mu}_t$ to be the mean of a finite sample.

For §5.2, we set $d = 5$ and $R = 5$. The matrix A was chosen as $-0.8\times$ a random PSD matrix. The vector b was set to be $2\times$ the all 1's vector. We used a time horizon of $T = 50$, and the noise on estimating the mean was $\sigma_{\text{err}} = 10^{-3}$. We did 5 trials per scenario.

For §5.3, we set $d = 5$, $R = 5$, and the mean estimation noise is still $\sigma_{\text{err}} = 10^{-3}$. We set $A = -0.8I$, b to be $2\times$ the all ones vector, and $\delta = 0.684$, and conducted 50 trials. In a small fraction of runs, the SPGD gradient estimate would explode, so we also clipped the gradient if its norm exceeded 10 by normalizing it to a unit vector.

For §5.4, we set $d = 2$ and $R = 3$. We had $\mu_0 = [2, 1]^\top$ and $\mu_1 = [1, 2]^\top$. The proportion of spammers was 0.5. We set $\alpha = -2$, the regularization strength to be 10^{-1}, and $\delta = 0.25$. The mean estimation noise was $\sigma_{\text{err}} = 10^{-3}$, and we conduct 50 trials.

C.1 Bottleneck

The long-term mean for θ is given by

$$\mu^*(\theta) = \frac{1}{1+\theta^\top \mu_0}\mu_0.$$

The long-term performative loss is then

$$\mathcal{L}^*(\theta) = \frac{-\theta^\top \mu_0}{1+\theta^\top \mu_0} + \frac{\lambda}{2}\|\theta\|^2.$$

If there is a long-term distribution, then the mean satisifes the fixed point equation $\mu = (1-\theta^\top \mu)\mu_0$. This equation implies that $\mu = c\mu_0$ for some scalar c. Substituting and solving the resulting equation, we see that The long-term mean for θ is given by

$$\mu^*(\theta) = \frac{1}{1+\theta^\top \mu_0}\mu_0.$$

To avoid the denominator blowing up, we would like to enforce some constraints on θ and μ. We accomplish this by setting $\Theta = \{\theta \in \mathbb{R}^d : \theta \geq 0 \text{ and } \|\theta\|_\infty \leq 1/\sqrt{d}\}$. If we also choose μ_0 so that $\|\mu_0\|_2 = 1$ and $\mu_0 \geq 0$, then we claim that $\|\mu_t\| \leq 1$ and $0 \leq \theta^\top \mu_t \leq 1$ for all t. It trivially holds for $t = 0$. At time $t+1$, we have:

$$\begin{aligned}\|\mu_{t+1}\|_2 &= |1-\theta^\top \mu_t|\|\mu_0\|_2 \\ &= |1-\theta^\top \mu_t|.\end{aligned}$$

Since $\theta^\top \mu_t \geq 0$, we trivially have $1-\theta^\top \mu_t \leq 1$. To see that it is also nonnegative:

$$\begin{aligned}1-\theta^\top \mu_t &\geq 1 - \|\theta\|_2\|\mu_t\|_2 \\ &\geq 1 - \sqrt{d}\frac{1}{\sqrt{d}} \cdot 1 = 0.\end{aligned}$$

It follows that $\|\mu_{t+1}\|_2 \leq 1$. Furthermore, since $\mu_0 \geq 0$ and $1-\theta^\top \mu_t \geq 0$, we have $\mu_{t+1} \geq 0$ and therefore $0 \leq \theta^\top \mu_{t+1} \leq 1$, completing the induction.

A simple calculation yields

$$\nabla \mathcal{L}^*(\theta) = \lambda\theta - \frac{\mu_0}{(1+\theta^\top \mu_0)^2}, \tag{97}$$

$$\nabla^2 \mathcal{L}^*(\theta) = \lambda I + \frac{2}{(1+\theta^\top \mu_0)^3}\mu_0\mu_0^\top. \tag{98}$$

If we assume that $\|\mu_0\|_2 = 1$, then (98) implies that $\mathcal{L}^*$ is convex precisely when $2/(1+\theta^\top \mu_0)^3 \geq -\lambda$ for all θ. Since $\mu_0, \theta \geq 0$, it follows that $\mathcal{L}^*$ is convex for any nonnegative regularization strength λ. Thus we should expect (approximate) gradient descent to find the minimizer for this problem.

D ADDITIONAL EXPERIMENTS

D.1 Extended Results for the Linear Experiment

Here we extend the results of Section 5.2 as the mean takes longer and longer to settle. For both of the following experiments, we kept the number of model deployments at $T = 50$. Figure 5 shows the performance of each algorithm at the same noise level as the previous experiment ($\sigma_{\text{err}} = 10^{-3}$). Figure 6 shows the results with no noise on the mean ($\sigma_{\text{err}} = 0$).

At the same noise level as the previous experiments, SPGD maintains its superior performance when it takes the distribution 128 steps to settle. However, as the number of steps required for the distribution to settle increases, the noise in mean estimation becomes larger than $\partial_2 m$ and SPGD can no longer get an accurate estimate of the long-term loss gradient. However, as the error on $\hat{\mu}_t$ decreases below the size of $\partial_2 m$, SPGD is able to form a good estimate of the long-term performative gradient even for extremely slowly adapting distributions, obtaining near-optimal performance even when the distribution takes 512 or even 2048 steps to settle to its long term value.

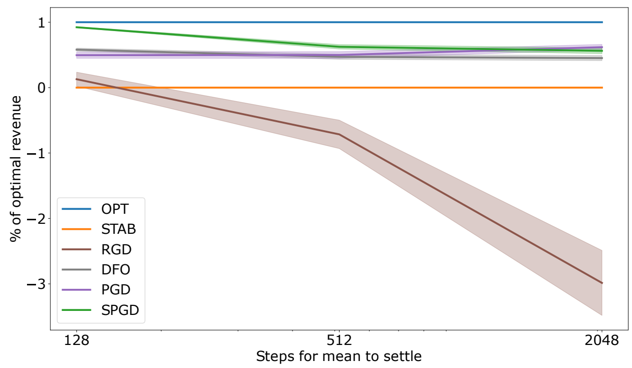

Figure 5: Additional results for the linear m experiment. SPGD retains its superior performance even when the distribution takes 128 steps to settle, but for slow enough dynamics, SPGD eventually degrades to the level of the other algorithms due to the noise in estimating μ_t.

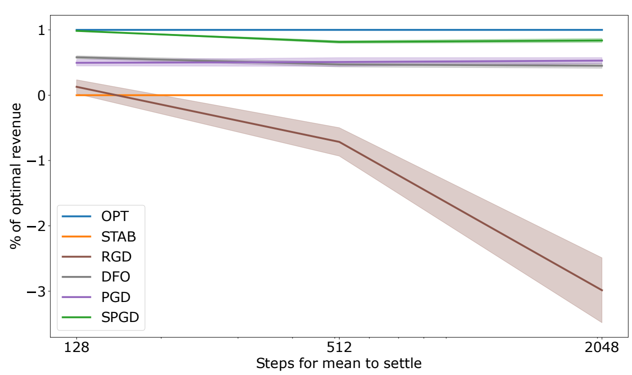

Figure 6: When there is no (or very little) noise in estimating μ_t, SPGD can still get an accurate estimate for the long-term gradient and manages to retain its superior performance even for extremely slowly adapting distributions.

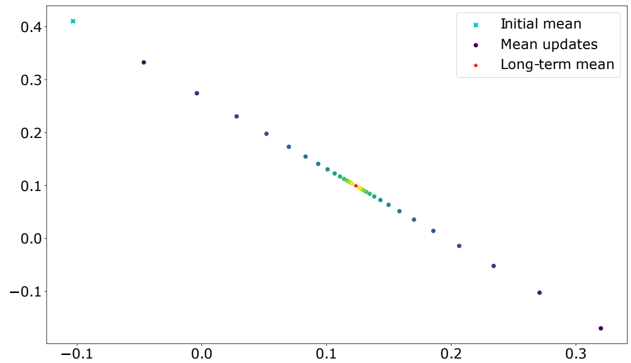

Figure 7: A demonstration of the oscillating distribution dynamics for this experiment.

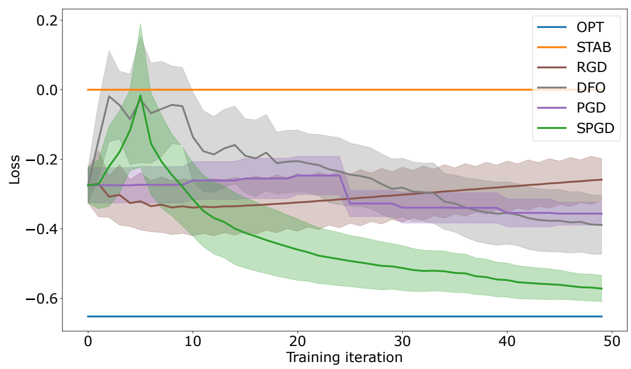

Figure 8: Performance of each of the algorithms for the oscillating distribution dynamics shown in Figure 7. SPGD still outperforms the other algorithms and finds a near-optimal point.

D.2 Dynamics with Oscillations

We consider another experiment where the distribution parameters do not converge monotonically to their long-term values. We still use the point loss $\ell(z, \theta) = -z^\top \theta$, but we take

$$m(\theta, \mu) = \delta(A\theta + b) - (1 - \delta)\mu$$

for some fixed $0 < \delta < 1$. In this case, $m^{(k)}$ oscillates around the long-term value of $\mu^*(\theta) = \frac{\delta}{2-\delta}(A\theta + b)$. This oscillation can be seen in Figure 7 in dimension $d = 2$. Consecutive updates of μ oscillate back and forth on either side of the long-term value.

In spite of the oscillations present in the dynamics, by choosing a *fixed* base point for the finite difference approximations used to estimate the long-term derivatives, SPGD still performs well in this setting. Figure 8 shows the results for $\delta = 0.134$. (Roughly speaking, this corresponds to a situation where it takes 32 steps for the effect of the initial distribution to decay.)