Throughout the following proofs, we use $\mathcal{O}(\cdot)$ to denote the leading order behavior of various quantities as T (the total number of steps taken by the method) becomes large and δ (the size of the error in the estimate for f) becomes small.

A. Bounding the error of $\widehat{f'}$

Before we prove Lemma 1 (bounding the error of the full performative gradient), we must first bound the error of our approximation to f'. Let $f_t = f(\theta_t)$, $f'_t = f'(\theta_t)$, and define $\hat{f}_t$ and $\widehat{f'_t}$ similarly.

Lemma 6. *Under the assumptions of Section 4, we have* $|\widehat{f'_t} - f'_t| = \mathcal{O}\left(\frac{\delta}{g}\frac{1}{\eta} + MG\eta\right)$.

Proof. By definition, we have

$$\widehat{f'_t} = \frac{\hat{f}_{t+1} - \hat{f}_t}{\theta_{t+1} - \theta_t} = \frac{f_{t+1} - f_t}{\theta_{t+1} - \theta_t} + \frac{\varepsilon_{t+1} - \varepsilon_t}{\theta_{t+1} - \theta_t}. \tag{8}$$

By Taylor's theorem, we have

$$f_{t+1} = f_t + f'_t \cdot (\theta_{t+1} - \theta_t) + \frac{1}{2} f''(\xi)(\theta_{t+1} - \theta_t)^2 \implies f'_t = \frac{f_{t+1} - f_t}{\theta_{t+1} - \theta_t} + \frac{1}{2} f''(\xi)(\theta_{t+1} - \theta_t), \tag{9}$$

where ξ is some number between θ_t and θ_{t+1}. Next, note that since $\theta_{t+1} = \theta_t - \eta \hat{\nabla}\mathcal{L}_t$ and $g \leq |\hat{\nabla}\mathcal{L}_t| \leq G$, we have $\eta g \leq |\theta_{t+1} - \theta_t| \leq \eta G$. Using this fact and combining equations (8) and (9), we find that

$$\begin{aligned} |\widehat{f'_t} - f'_t| &\leq \frac{|\varepsilon_{t+1}| + |\varepsilon_t|}{\eta g} + \frac{1}{2}|f''(\xi)|\eta G \\ &\leq \frac{2\delta}{g}\frac{1}{\eta} + \frac{MG}{2}\eta \end{aligned}$$

where we have also used the assumption that $|f''(\xi)| \leq M$. This is the desired bound. □

B. Proof of Lemma 1

Proof. We write $|\hat{\nabla}\mathcal{L}_t - \nabla\mathcal{L}_t| \leq |\hat{\nabla}_1\mathcal{L}_t - \nabla_1\mathcal{L}_t| + |\hat{\nabla}_2\mathcal{L} - \nabla_2\mathcal{L}_t|$ and bound each term on the right-hand side separately. We begin by bounding the error on $\nabla_1\mathcal{L}$. We have

$$\begin{aligned} |\hat{\nabla}_1\mathcal{L}_t - \nabla_1\mathcal{L}_t| &\leq \int |\nabla\ell(z;\theta)||p(z;\hat{f}_t) - p(z;f_t)|\,dz \\ &\leq \ell_{\max}\int |p(z;\hat{f}_t) - p(z;f_t)|\,dz \\ &\leq \ell_{\max}\Bigg(\underbrace{\int_{|z-f_t|\leq R} |p(z;\hat{f}_t) - p(z;f_t)|\,dz}_{\text{(A)}} + \underbrace{\int_{|z-f_t|>R} |p(z;\hat{f}_t)|\,dz}_{\text{(B)}} + \underbrace{\int_{|z-f_t|>R} |p(z;f_t)|\,dz}_{\text{(C)}}\Bigg), \end{aligned} \tag{10}$$

where for simplicity we assume that $\ell_{\max} \geq |\nabla\ell(z;\theta)|$ is also an upper bound on the derivative of the point loss, and for any $R > 0$.

To bound (A), we bound the Lipschitz constant of p in its second argument. It suffices to bound $\partial_2 p$. Observe that

$$\partial_2 p(z;w) = c(z-w)e^{-\frac{1}{2\sigma^2}(z-w)^2}. \tag{11}$$

Letting $x = z - w$ and $\alpha = \frac{1}{2\sigma^2}$, we want to bound the maximum of $xe^{-\alpha x^2}$. Taking the derivative with respect to x, this has critical points at $x = \pm\frac{1}{\sqrt{2\alpha}}$. Since $|\partial_2 p(z;w)| \to 0$ as $w \to \pm\infty$ for any z, these critical points are global maxima for $|\partial_2 p|$. Thus $\max|\partial_2 p| = \mathcal{O}(1)$ and p is $\mathcal{O}(1)$-Lipschitz in its second argument. It follows that

$$\text{(A)} \leq \int c|\hat{f}_t - f_t|\,dz = \mathcal{O}(R\delta).$$

To bound (B), oberserve that

$$\begin{aligned}(\mathrm{B}) &\leq \int_{|z-\hat{f}_t|+|\hat{f}_t-f_t|>R} |p(z;\hat{f}_t)|\,dz \\ &\leq \int_{|z-\hat{f}_t|>R-\delta} |p(z;\hat{f}_t)|\,dz \\ &= \mathbb{P}_{\mathcal{N}(\hat{f}_t,\sigma^2)}(|z-\hat{f}_t|>R-\delta) \\ &\leq 2e^{-(R-\delta)^2/2\sigma^2}\end{aligned}$$

for any $R \geq \delta$. A similar calculation shows that (C) $\leq 2e^{-R^2/2\sigma^2} \leq 2e^{-(R-\delta)^2/2\sigma^2}$ for $R \geq \delta$. Thus

$$(\mathrm{A})+(\mathrm{B})+(\mathrm{C}) = \mathcal{O}\left(R\delta + \exp\left(-\frac{(R-\delta)^2}{2\sigma^2}\right)\right)$$

for any $R \geq \delta$. Setting $R = \delta + \sqrt{2\sigma^2 \log \frac{1}{\delta}}$ and substituting our bound back into (10), we obtain

$$|\hat{\nabla}_1\mathcal{L}_t - \nabla_1\mathcal{L}_t| = \mathcal{O}\left(\ell_{\max}\left(\delta\sqrt{\log\frac{1}{\delta}}\right)\right). \tag{12}$$

Next we bound the error $|\hat{\nabla}_2\mathcal{L}_t - \nabla_2\mathcal{L}_t|$. We have

$$\begin{aligned}|\hat{\nabla}_2\mathcal{L}_t - \nabla_2\mathcal{L}_t| &= \left|\int \ell(z;\theta_t)\partial_2 p(z;\hat{f}_t)\widehat{f}_t'\,dz - \int \ell(z;\theta_t)\partial_2 p(z;f_t)f_t'\,dz\right| \\ &\leq \underbrace{\int |\ell(z;\theta_t)||\partial_2 p(z;\hat{f}_t)||\widehat{f}_t' - f_t'|\,dz}_{\text{(I)}} + \underbrace{\int |\ell(z;\theta_t)||\partial_2 p(z;\hat{f}_t) - \partial_2 p(z;f_t)||f_t'|\,dz}_{\text{(II)}}.\end{aligned} \tag{13}$$

We proceed to bound the terms (I) and (II) separately.

The bound for (I) is straightforward. Recall that $|\ell(z;\theta_t)| \leq \ell_{\max}$ and $\widehat{f}_t'$ and f_t' are independent of z, so we have

$$(\mathrm{I}) \leq \ell_{\max}|\widehat{f}_t' - f_t'| \int |\partial_2 p(z;\hat{f}_t)|\,dz.$$

Since $p(z;\hat{f}_t)$ is the pdf for a Gaussian with mean $\hat{f}_t$ and variance σ^2, a standard computation reveals that $\int |\partial_2 p(z;\hat{f}_t)|\,dz = \sqrt{\frac{2}{\pi\sigma^2}} = \mathcal{O}(1)$. Using the bound on $|\widehat{f}_t' - f_t'|$ from Lemma 6, we have

$$(\mathrm{I}) = \mathcal{O}\left(\ell_{\max}\left(MG\eta + \frac{\delta}{g}\frac{1}{\eta}\right)\right). \tag{14}$$

Next, we bound (II). First, since $|\ell(z;\theta_t)| \leq \ell_{\max}$ and $|f_t'| \leq F$, we have

$$(\mathrm{II}) \leq \ell_{\max}F\int |\partial_2 p(z;\hat{f}_t) - \partial_2 p(z;f_t)|\,dz \tag{15}$$

so it suffices to bound the integrand in (15).

For any $R \geq \delta$, we have

$$\int |\partial_2 p(z;\hat{f}_t) - \partial_2 p(z;f_t)|\,dz = \underbrace{\int_{|z-f_t|\leq R} |\partial_2 p(z;\hat{f}_t) - \partial_2 p(z;f_t)|\,dz}_{\text{(i)}} + \underbrace{\int_{|z-f_t|>R} |\partial_2 p(z;\hat{f}_t) - \partial_2 p(z;f_t)|\,dz}_{\text{(ii)}}.$$

To bound (i), it suffices to bound the Lipschitz constant of $\partial_2 p(z;w)$ in the second variable (if one exists). We can do this by bounding $|\partial_2^2 p|$. A direct computation shows that

$$\partial_2^2 p(z;w) = \frac{1}{\sigma^2\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(z-w)^2}\left(\frac{1}{\sigma^2}(z-w)^2 - 1\right). \tag{16}$$

Let $\alpha = \sigma^{-2}$ and $x = (z-w)^2$. Bounding (16) is equivalent to upper bounding an expression of the form $e^{-\frac{\alpha}{2}x}(\alpha x - 1)$ over $x \geq 0$. Taking a derivative with respect to x shows that the only critical point is at $x = 2e^{-3/2}$; the only other point to check is the boundary point $x = 0$. Checking both of these manually shows that the absolute value is maximized at $x = 0$, and we obtain the bound

$$|\partial_2^2 p(z;w)| \leq \frac{1}{\sigma^2\sqrt{2\pi\sigma^2}} = \mathcal{O}(1),$$

i.e. $\partial_2 p(z;w)$ is $\mathcal{O}(1)$-Lipschitz in w. Applying this fact to (i), we have

$$\text{(i)} \leq \int_{|z-f_t|\leq R} c|\hat{f}_t - f_t|\, dz = c\int_{|z-f_t|\leq R} |\varepsilon_t|\, dz = \mathcal{O}(R\delta). \tag{17}$$

Next we turn our attention to (ii). We have

$$\begin{aligned}
\text{(ii)} &\leq \int_{|z-f_t|>R} |\partial_2 p(z;\hat{f}_t)|\, dz + \int_{|z-f_t|>R} |\partial_2 p(z;f_t)|\, dz \\
&\leq \int_{|z-(f_t+\varepsilon_t)|>R-|\varepsilon_t|} |\partial_2 p(z;\hat{f}_t)|\, dz + \int_{|z-f_t|>R} |\partial_2 p(z;f_t)|\, dz \\
&\leq \int_{|z-\hat{f}_t|>R-\delta} |\partial_2 p(z;\hat{f}_t)|\, dz + \int_{|z-f_t|>R} |\partial_2 p(z;f_t)|\, dz. \qquad (18)
\end{aligned}$$

These inequalities follow from several applications of the triangle inequality and the bound $|\varepsilon_t| \leq \delta$. Now since $p(z;w)$ is a Gaussian pdf, we have $\partial_2 p(z;w) = \frac{1}{\sigma^2}(z-w)p(z;w)$, and therefore

$$\begin{aligned}
\int_{|z-w|>r} |\partial_2 p(z;w)|\, dz &= \int_{|z-w|>r} \frac{1}{\sigma^2}|z-w|p(z;w)\, dz \\
&= \sigma^{-2}\mathbb{E}_{\mathcal{N}(w,\sigma^2)}\left[\mathbb{1}\{|z-w| \geq r|\}|z-w|\right] \\
&\leq \sigma^{-2}\sqrt{\mathbb{E}[\mathbb{1}\{|z-w| \geq r\}^2]\mathbb{E}[|z-w|^2]} \qquad (19) \\
&= \sigma^{-1}\sqrt{\mathbb{P}(|z-w| \geq r)} \\
&\leq \sqrt{2\sigma^{-1}}e^{-\frac{r^2}{4\sigma^2}}, \qquad (20)
\end{aligned}$$

where (19) follows from the Cauchy-Schwarz inequality and (20) follows from a standard Gaussian tail bound. Applying (20) to (18), we obtain

$$\begin{aligned}
\text{(ii)} &\leq c\left(\exp\left(-\frac{(R-\delta)^2}{4\sigma^2}\right) + \exp\left(-\frac{R^2}{4\sigma^2}\right)\right) \\
&= \mathcal{O}\left(\exp\left(-\frac{(R-\delta)^2}{4\sigma^2}\right)\right) \qquad (21)
\end{aligned}$$

for any $R \geq \delta$. Combining the bound (17) on (i) and (21) on (ii) with (15), we have

$$\text{(II)} = \mathcal{O}\left(\ell_{\max}F\left[R\delta + \exp\left\{-\frac{(R-\delta)^2}{4\sigma^2}\right\}\right]\right). \tag{22}$$

If we take $R = \delta + \sqrt{4\sigma^2 \log(1/\delta)}$ and substitute into (22), we obtain

$$(\mathrm{II}) = \mathcal{O}\left(\ell_{\max} F\delta\sqrt{\log(1/\delta)}\right). \tag{23}$$

We now substitute our bounds on (I) and (II) into (13), which yields

$$|\nabla_2\mathcal{L} - \hat{\nabla}_2\mathcal{L}| \leq \mathcal{O}\left(\ell_{\max}\left[MG\eta + \frac{\delta}{g}\frac{1}{\eta} + F\delta\sqrt{\log(1/\delta)}\right]\right). \tag{24}$$

To conlude, observe that the bound on the error of $\nabla_1\mathcal{L}_t$ in (12) can be completely absorbed into (24), and we obtain the desired result. □

C. Proof of Theorem 2

Proof. To simplify notation, we will let $L = L_{\mathrm{Lip}}$; this should not be confused with the decoupled performative loss function $L(\theta_1, \theta_2)$ defined in Section 2. Let $\mathcal{L}_t = \mathcal{L}(\theta_t)$ and let $E_t = \hat{\nabla}\mathcal{L}_t - \nabla\mathcal{L}_t$. Since $\mathcal{L}$ is L-smooth and convex, we have the standard inequality

$$\mathcal{L}_{t+1} \leq \mathcal{L}_t + \nabla\mathcal{L}_t \cdot (\theta_{t+1} - \theta_t) + \frac{L}{2}|\theta_{t+1} - \theta_t|^2. \tag{25}$$

Since $\theta_{t+1} - \theta_t = \eta\hat{\nabla}\mathcal{L}_t$, we can rewrite (25):

$$\mathcal{L}_{t+1} \leq \mathcal{L}_t + \eta(|\nabla\mathcal{L}_t||E_t| - |\nabla\mathcal{L}_t|^2) + \eta^2 L(|\nabla\mathcal{L}_t|^2 + |E_t|^2) \tag{26}$$

Rearranging and using the fact that $|\nabla\mathcal{L}_t| \leq G$, we have

$$(\eta - \eta^2 L)|\nabla\mathcal{L}_t|^2 \leq \mathcal{L}_t - \mathcal{L}_{t+1} + \eta G|E_t| + \eta^2 L|E_t|^2. \tag{27}$$

If we sum both sides of (27) from $t = 1$ to T, we find that

$$T\min_{1\leq t\leq T}|\nabla\mathcal{L}_t|^2 \leq \sum_{t=1}^{T}|\nabla\mathcal{L}_t|^2 \leq \frac{\mathcal{L}_1 - \mathcal{L}_{T+1} + \eta G\sum_{t=1}^{T}|E_t| + \eta^2 L\sum_{t=1}^{T}|E_t|^2}{\eta - L\eta^2}. \tag{28}$$

Note that with $\eta = \sqrt{\frac{1}{MG^2T} + \frac{\delta}{MGg}}$ as specified by the theorem, we have $\eta^2 = o(\eta)$. Furthermore, by Lemma 1, we have

$$|E_t| = \mathcal{O}\left(\ell_{\max}\sqrt{\frac{M}{T} + \frac{MG\delta}{g}}\right) \equiv \mathbf{E}. \tag{29}$$

(In obtaining the above bound, we have assumed WLOG that $G \geq 1$.) Note that since $\mathbf{E} = o(1)$, we have $\mathbf{E}^2 = o(\mathbf{E})$. Lastly, since $\mathcal{L}_t = \mathbb{E}_{p(z;\theta_t)}[\ell(z;\theta_t)]$ we have $|\mathcal{L}_t| \leq \ell_{\max}$ for all t. Applying these facts to (28), we have

$$\begin{aligned}\min_{1\leq t\leq T}|\nabla\mathcal{L}_t|^2 &= \mathcal{O}\left(\frac{\ell_{\max} + \eta GT\mathbf{E} + \eta^2 LT\mathbf{E}^2}{T\eta}\right)\\ &= \mathcal{O}\left(\frac{\ell_{\max}}{T\eta} + G\ell_{\max}\left[MG\eta + \frac{\delta}{g}\frac{1}{\eta}\right]\right)\\ &= \mathcal{O}\left(\ell_{\max}\sqrt{\frac{MG^2}{T} + \frac{MG^3\delta}{g}}\right)\end{aligned} \tag{30}$$

where the last equation follows from our choice of η.

Lastly, recall that our bound on $|E_t|$ required that $|\hat{\nabla}\mathcal{L}_t| \geq g$ for all $1 \leq t \leq T$. If at any point we have $|\hat{\nabla}\mathcal{L}_t| < g$, then we can terminate and return this iterate. But then we have

$$|\nabla\mathcal{L}_t|^2 \leq 2|\hat{\nabla}\mathcal{L}_t|^2 + 2|E_t|^2 = \mathcal{O}(g^2 + \mathbf{E}^2).$$

Note that $\mathbf{E}^2 = \mathcal{O}\left(\ell_{\max}^2 \left(\frac{M}{T} + \frac{MG\delta}{g}\right)\right)$, which implies that

$$|\nabla \mathcal{L}_t|^2 = \mathcal{O}\left(g^2 + \ell_{\max}^2 \left(\frac{M}{T} + \frac{MG\delta}{g}\right)\right). \tag{31}$$

We can guarantee that PerfGD reaches at least the max of the two bounds (30) and (31), yielding the desired result. □

We remark that, for a given accuracy level δ, we should take a time horizon $T \propto \delta^{-1}$. Increasing T beyond this point will not cause the error bound from Theorem 2 to decay any further.

C.1. Proof of corollaries

Proof of Corollary 4. Suppose $n_t \geq n$ samples are collected at time t. Then the error $\hat{f}(\theta_t) - f(\theta_t)$ follows a centered Gaussian distribution with variance $\sigma^2/n_t \leq \sigma^2/n$. It follows that

$$\begin{aligned}\mathbb{P}(|\hat{f}(\theta_t) - f(\theta_t)| > \varepsilon) &\leq \mathbb{P}_{z\sim\mathcal{N}(0,\sigma^2/n)}(|z| > \varepsilon) \\ &\leq 2e^{-n\varepsilon^2/2\sigma^2} \\ &\leq \delta/T\end{aligned}$$

where the second inequality follows from a standard Gaussian tail bound and the third follows from the lower bound on n. Taking a union bound over all T steps of the algorithm, we see that $|\hat{f}_t - f_t| < \varepsilon$ for all t with probability at least $1 - \gamma$. □

D. Proofs for higher dimensions

The analysis used to prove Theorem 2 applies in general dimensions when we replace all absolute value signs with the Euclidean norm, provided that we have a bound on the error $E_t = \hat{\nabla}\mathcal{L}_t - \nabla\mathcal{L}_t$. To bound E_t, we will proceed as before, first bounding the error on the estimate for $df/d\theta$ and then translating this to a bound on the error of the full performative gradient.

Throughout this section, when applied to *vectors*, $\|\cdot\|$ will always denote the Euclidean norm. When applied to *matrices*, $\|\cdot\|_2$ denotes the matrix 2-norm and $\|\cdot\|_F$ denotes the Frobenius norm. For a PSD matrix Σ, we will also sometimes write $\lambda_{\text{max}}(\Sigma)$ and $\lambda_{\text{min}}(\Sigma)$ for the maximum and minimum eigenvalues of Σ, respectively.

Lemma 7. *When $H \geq d$ and $\Delta\theta$ has full row rank d, we have* $\left\|\widehat{\frac{df_t}{d\theta}} - \frac{df_t}{d\theta}\right\|_2 \leq \left(\frac{MdG^2H^3}{2}\eta + 2\delta\sqrt{H}\frac{1}{\eta}\right) g^{-1}$

Proof. Observe that, for each $1 \leq i \leq H$ and for each $1 \leq j \leq d$, we have

$$f_j(\theta_{t-i}) - f_j(\theta_t) = \nabla f_j(\theta_t)^\top(\theta_{t-i} - \theta_t) + \frac{1}{2}(\theta_{t-i} - \theta_t)^\top \nabla^2 f_j(\xi_{ij})(\theta_{t-i} - \theta_t),$$

where $\xi_{ij} \in [\theta_{t-i}, \theta_t]$ lies on the line segment joining θ_{t-i} and θ_t. This is a consequence of Taylor's theorem. If we assume that $\|\nabla^2 f_j\|_2 \leq M$ is bounded, then we have

$$f(\theta_{t-i}) - f(\theta_t) = \left.\frac{df}{d\theta}\right|_{\theta=\theta_t}(\theta_{t-i} - \theta_t) + \text{err}_i^{\text{Taylor}}$$

and $\|\text{err}_i^{\text{Taylor}}\|_2 \leq \frac{Md}{2}\|\theta_{t-i} - \theta_t\|^2$. Denote $\frac{df_t}{d\theta} = \left.\frac{df}{d\theta}\right|_{\theta=\theta_t}$. Observe that

$$
\begin{aligned}
\Delta f &= \begin{bmatrix} | & & | \\ \hat{f}_{t-1} - \hat{f}_t & \cdots & \hat{f}_{t-H} - \hat{f}_t \\ | & & | \end{bmatrix} \\
&= \begin{bmatrix} | & & | \\ \frac{df_t}{d\theta}(\theta_{t-1}-\theta_t) + \mathrm{err}_1^{\text{Taylor}} + \varepsilon_{t-1} - \varepsilon_t & \cdots & \frac{df_t}{d\theta}(\theta_{t-H}-\theta_t) + \mathrm{err}_H^{\text{Taylor}} + \varepsilon_{t-H} - \varepsilon_t \\ | & & | \end{bmatrix} \\
&= \frac{df_t}{d\theta}\Delta\theta + \underbrace{\begin{bmatrix} | & & | \\ \mathrm{err}_1^{\text{Taylor}} & \cdots & \mathrm{err}_H^{\text{Taylor}} \\ | & & | \end{bmatrix}}_{\mathrm{err}^{\text{Taylor}}} + \underbrace{\begin{bmatrix} | & & | \\ \varepsilon_{t-1} - \varepsilon_t & \cdots & \varepsilon_{t-H} - \varepsilon_t h \\ | & & | \end{bmatrix}}_{\varepsilon} \\
\implies \frac{\widehat{df_t}}{d\theta} &= \frac{df_t}{d\theta} + (\mathrm{err}^{\text{Taylor}} + \varepsilon)(\Delta\theta)^{\dagger}.
\end{aligned}
$$

This final equation holds because $\Delta\theta$ was assumed to have full row rank, therefore $(\Delta\theta)^{\dagger}$ is the right inverse of $\Delta\theta$. It follows that

$$
\begin{aligned}
\left\| \frac{\widehat{df_t}}{d\theta} - \frac{df_t}{d\theta} \right\|_2 &\le (\|\mathrm{err}^{\text{Taylor}}\|_F + \|\varepsilon\|_F)\|\Delta\theta^{\dagger}\|_2 \\
&\le \left(\sqrt{\sum_{i=1}^{H} \|\mathrm{err}_i^{\text{Taylor}}\|^2} + \sqrt{\sum_{i=1}^{H} \|\varepsilon_{t-i} - \varepsilon_t\|^2} \right) \sigma_{\text{min}}(\Delta\theta)^{-1} \\
&\le \left(\frac{Md}{2} \sum_{i=1}^{H} \|\theta_{t-i} - \theta_t\|^2 + 2\delta\sqrt{H} \right) (g\eta)^{-1} \\
&\le \left(\frac{MdG^2H^3}{2}\eta + 2\delta\sqrt{H}\frac{1}{\eta} \right) g^{-1}.
\end{aligned}
$$

□

Before bounding $\|\hat{\nabla}\mathcal{L}_t - \nabla\mathcal{L}_t\|$, we first need some regularity results for multivariate Gaussian distributions. We have not attempted to optimize our bounds with respect to the dependence on dimension or other problem-dependent quantities (e.g. properties of the covariance matrix, etc.). Our goal is to show that the performance of PerfGD in moderately-sized data dimension (i.e. $1 < d \ll \infty$) is qualitatively similar to its performance in 1D.

Lemma 8. *Let $p(z;\mu) = c\exp\{-\frac{1}{2}(z-\mu)^{\top}\Sigma^{-1}(z-\mu)\}$ be the d-dimensional normal pdf with (fixed) covariance Σ and mean μ. If the normalizing constant $c = (2\pi)^{-d/2}(\det\Sigma)^{-1/2} = \mathcal{O}(1)$, then p is $\mathcal{O}(1)$-Lipschitz in μ.*

Proof. It suffices to bound $\nabla_{\mu}p(z;\mu)$. We have

$$
\begin{aligned}
\|\nabla_{\mu}p(z;\mu)\| &= \left\| -c\Sigma^{-1}(\mu - z)\exp\left\{ -\frac{1}{2}(\mu - z)^{\top}\Sigma^{-1}(\mu - z) \right\} \right\| \\
&\le c\lambda_{\text{min}}(\Sigma)^{-1}\|\mu - z\| \exp\left\{ -\frac{1}{2}\lambda_{\text{max}}(\Sigma)^{-1}\|\mu - z\|^2 \right\} \\
&= Cxe^{-\alpha x^2}, \qquad (32)
\end{aligned}
$$

where $x = \|\mu - z\|$ and C and α collect constant terms. By the analysis of (11), it follows that (32) is also bounded by a constant, and thus p is $\mathcal{O}(1)$-Lipschitz in the mean μ. (Note: Here we implicitly assume that problem-specific parameters such as $(\det\Sigma)^{-1/2}$, $\lambda_{\text{min}}(\Sigma)^{-1}$, and $\lambda_{\text{max}}(\Sigma)$ are all $\mathcal{O}(1)$.) □

Lemma 9. *If $Z \sim \mathcal{N}(\mu, \Sigma)$ is a d-dimensional Gaussian vector, then $\mathbb{P}(\|Z - \mu\| \ge R) \le 2d\exp\{-\frac{R^2}{2d\lambda_{\max}(\Sigma)}\}$.*

Proof. A pigeonhole principle-esque argument shows that $\mathbb{P}(\|Z-\mu\| \geq R) \leq \mathbb{P}(|Z_i - \mu_i| \geq R/\sqrt{d}$ for some $i)$, where Z_i, μ_i denote the i-th coordinate of Z and μ respectively. Note that each Z_i is normally distributed with mean μ_i and variance at most $\lambda_{\text{max}}(\Sigma)$. By a union bound, it follows that

$$\begin{aligned}\mathbb{P}(\|Z-\mu\| \geq R) &\leq \sum_{i=1}^{d} \mathbb{P}(|Z_i - \mu_i| \geq R/\sqrt{d}) \\ &\leq d\mathbb{P}_{z\sim\mathcal{N}(0,\lambda_{\text{max}}(\Sigma))}(|z| \geq R/\sqrt{d}) \\ &\leq 2d\exp\left\{\frac{-R^2}{2d\lambda_{\text{max}}(\Sigma)}\right\}\end{aligned}$$

where the final inequality follows from a standard one-dimensional Gaussian tail bound. □

Lemma 10. $\int \|\nabla_\mu p(z;\mu)\| \, dz \leq \|\Sigma^{-1/2}\|_2\sqrt{d}$.

Proof.

$$\begin{aligned}\int \|\nabla_\mu p(z;\mu)\| \, dz &= \int \|\Sigma^{-1}(z-\mu)\| p(z;\mu) \, dz \\ &\leq \|\Sigma^{-1/2}\|_2 \int \|\Sigma^{-1/2}(z-\mu)\| p(z;\mu) \, dz \\ &= \|\Sigma^{-1/2}\|_2 \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[\|z\|] \\ &\leq \|\Sigma^{-1/2}\|_2\sqrt{d}.\end{aligned}$$

□

Lemma 11. *$\nabla_\mu p(z;\mu)$ is $\mathcal{O}(1)$-Lipschitz in μ.*

Proof. It suffices to show that $\|\nabla_\mu^2 p\|_2 = \mathcal{O}(1)$. A simple calculation yields

$$\nabla_\mu^2 p = \Sigma^{-1}(pI + \Sigma^{-1}p(\mu - z)(\mu - z)^\top).$$

From this it follows that

$$\|\nabla_\mu^2 p\|_2 \leq \|\Sigma^{-1}\|_2((\sup p) + \|\Sigma^{-1}\|_2 p(z;\mu)\|\mu - z\|^2).$$

All of the quantities in this expression are $\mathcal{O}(1)$ by assumption except for $p(z;\mu)\|\mu - z\|^2$, which has the form

$$\begin{aligned}p(z;\mu)\|\mu - z\|^2 &= c\|\mu - z\|^2 \exp\left\{-\frac{1}{2}(\mu - z)^\top \Sigma^{-1}(\mu - z)^\top\right\} \\ &\leq c\|\mu - z\|^2 \exp\left\{-\frac{1}{2\lambda_{\text{max}}(\Sigma)}\|\mu - z\|^2\right\} \\ &= cx^2 e^{-\alpha x^2},\end{aligned}$$

where again $x = \|\mu - z\|$ and α collects the relevant constants. By the analysis from (16), this final quantity is bounded by a constant. It follows that $\|\nabla_\mu^2 p\|_2 = \mathcal{O}(1)$ as desired. □

Lemma 12. $\|\hat{\nabla}\mathcal{L}_t - \nabla\mathcal{L}_t\| = \mathcal{O}\left(\ell_{\text{max}}\left[\sqrt{d}g^{-1}(MdG^2H^3\eta + \delta\sqrt{H}\frac{1}{\eta}) + F\delta\sqrt{d\log\frac{d}{\delta}}\right]\right)$.

Proof. We plug the results of the previous four lemmas into the proof of Lemma 1. The error on $\nabla_1\mathcal{L}_t$ is bounded by

$$\|\hat{\nabla}_1\mathcal{L}_t - \nabla_1\mathcal{L}_t\| \leq \ell_{\text{max}}\Bigg(\underbrace{\int_{\|z-f_t\|\leq R} |p(z;\hat{f}_t) - p(z;f_t)| \, dz}_{\text{(A)}} + \underbrace{\int_{\|z-f_t\|>R} |p(z;\hat{f}_t)| \, dz}_{\text{(B)}} + \underbrace{\int_{\|z-f_t\|>R} |p(z;f_t)| \, dz}_{\text{(C)}}\Bigg), \tag{33}$$

Since p is $\mathcal{O}(1)$-Lipschitz in its second argument by Lemma 8, (A) is bounded by $\mathcal{O}(R\delta)$. For (B), when $R \geq \delta$ we have

$$\begin{aligned}\text{(B)} &\leq \int_{\|z-\hat{f}_t\|>R-\delta} |p(z;\hat{f}_t)|\, dz \\ &\leq 2d\exp\{-(R-\delta)^2/2d\sigma^2\},\end{aligned}$$

where $\sigma^2 = \lambda_{\text{max}}(\Sigma)$ and the second inequality follows from Lemma 9. We similarly have (C) $\leq 2d\exp\{-(R-\delta)^2/2d\sigma^2\}$. Taking $R = \delta + \sqrt{2d\sigma^2 \log \frac{d}{\delta}}$, we have

$$\|\hat{\nabla}_1\mathcal{L}_t - \nabla_1\mathcal{L}_t\| = \mathcal{O}\left(\delta\sqrt{d\log\frac{d}{\delta}}\right). \tag{34}$$

Next we bound $\|\hat{\nabla}_2\mathcal{L}_t - \nabla_2\mathcal{L}_t\|$. As before, we have

$$\|\hat{\nabla}_2\mathcal{L}_t - \nabla_2\mathcal{L}_t\| \leq \underbrace{\ell_{\text{max}}\left\|\widehat{\frac{df_t}{d\theta}}^\top - \frac{df_t}{d\theta}^\top\right\|_2 \int \|\nabla_2 p(z;\hat{f}_t)\|\, dz}_{\text{(I)}} + \underbrace{\ell_{\text{max}}F\int\|\nabla_2 p(z;\hat{f}_t) - \nabla_2 p(z;f_t)\|\, dz}_{\text{(II)}}.$$

By Lemma 10, the integral in (I) is bounded by $\mathcal{O}(\sqrt{d})$. Plugging in the bound from Lemma 7 and noting that $\left\|\widehat{\frac{df_t}{d\theta}}^\top - \frac{df_t}{d\theta}^\top\right\|_2 = \left\|\widehat{\frac{df_t}{d\theta}} - \frac{df_t}{d\theta}\right\|_2$, we have

$$\text{(I)} = \mathcal{O}\left(\ell_{\text{max}}\sqrt{d}g^{-1}\left(MdG^2H^3\eta + \delta\sqrt{H}\frac{1}{\eta}\right)\right).$$

Bounding (II) is similar to before. For $R \geq \delta$, we write

$$\int\|\nabla_2 p(z;\hat{f}_t) - \nabla_2 p(z;f_t)\|\, dz = \underbrace{\int_{\|z-f_t\|\leq R}\|\nabla_2 p(z;\hat{f}_t) - \nabla_2 p(z;f_t)\|\, dz}_{\text{(i)}} + \underbrace{\int_{\|z-f_t\|>R}\|\nabla_2 p(z;\hat{f}_t) - \nabla_2 p(z;f_t)\|\, dz}_{\text{(ii)}}.$$

By Lemma 11, $\nabla_2 p$ is $\mathcal{O}(1)$-Lipschitz in the mean parameter, so (i) $= \mathcal{O}(R\delta)$.

To bound (ii), we use a similar strategy as from the proof of Lemma 1. We first observe that

$$\text{(ii)} \leq \int_{\|z-\hat{f}_t\|>R-\delta}\|\nabla_2 p(z;\hat{f}_t)\|\, dz + \int_{\|z-f_t\|>R}\|\nabla_2 p(z;f_t)\|\, dz. \tag{35}$$

We then have

$$\begin{aligned}\int_{\|z-\hat{f}_t\|>R-\delta}\|\nabla_2 p(z;\hat{f}_t)\|\, dz &= \int_{\|z-\hat{f}_t\|>R-\delta}\|\Sigma^{-1}(z-\hat{f}_t)\|p(z;\hat{f}_t)\, dz \\ &= \mathbb{E}_{z\sim\mathcal{N}(\hat{f}_t,\Sigma)}[\mathbb{1}\{\|z-\hat{f}_t\| > R-\delta\}\|\Sigma^{-1}(z-f_t)\|] \\ &\leq \sqrt{\mathbb{P}(\|z-\hat{f}_t\| > R-\delta)\cdot\|\Sigma^{-1/2}\|_2\mathbb{E}_{\mathcal{N}(\hat{f}_t,\Sigma)}[\|\Sigma^{-1/2}(z-\hat{f}_t)\|^2]} \\ &\leq \sqrt{2d\exp\left\{\frac{-(R-\delta)^2}{2d\lambda_{\text{max}}(\Sigma)}\right\}\cdot\|\Sigma^{-1/2}\|_2 d} \\ &= \mathcal{O}\left(d\exp\left\{\frac{-(R-\delta)^2}{4d\lambda_{\text{max}}(\Sigma)}\right\}\right).\end{aligned}$$

The same bound holds for the other integral in (35), so (ii) $= \mathcal{O}(d\exp\{\frac{-(R-\delta)^2}{4d\lambda_{\max}(\Sigma)}\})$. Combining this with the bound on (i), we have

$$\text{(II)} = \mathcal{O}\left(\ell_{\max}F\left[R\delta + d\exp\left\{\frac{-(R-\delta)^2}{4d\lambda_{\max}(\Sigma)}\right\}\right]\right).$$

If we take again set $\sigma^2 = \lambda_{\max}(\Sigma)$ and take $R = \delta + \sqrt{4d\sigma^2\log\frac{d}{\delta}}$, then we finally arrive at

$$\text{(II)} = \mathcal{O}\left(\ell_{\max}F\delta\sqrt{d\log\frac{d}{\delta}}\right).$$

Combining the bounds on (I) and (II), we have

$$\|\hat{\nabla}_2\mathcal{L}_t - \nabla_2\mathcal{L}_t\| = \mathcal{O}\left(\ell_{\max}\left[\sqrt{d}g^{-1}\left(MdG^2H^3\eta + \delta\sqrt{H}\frac{1}{\eta}\right) + F\delta\sqrt{d\log\frac{d}{\delta}}\right]\right).$$

Note that once again, the error bound on $\nabla_1\mathcal{L}_t$ can be completely absorbed into this expression, yielding the desired result. □

Proof of Theorem 5. The proof of Theorem 2 from Appendix C applies with two slight modifications. First, all absolute values should be replaced with the Euclidean norm. Second, we use a different value for η. With this step size, we still have $\eta^2 = o(\eta)$ and $\mathbf{E}^2 = o(\mathbf{E})$ (recall that $\mathbf{E}$ is defined as the bound on the performative gradient error from Lemma 12), so the exact same logic as before gives

$$\min_{1\le t\le T}\|\nabla\mathcal{L}_t\|^2 = \mathcal{O}\left(\frac{\ell_{\max}}{T\eta} + G\mathbf{E}\right).$$

Plugging in the new value for η as well as the expression for $\mathbf{E}$ from Lemma 12 yields the desired bound. □

E. Convergence of PerfGD with stochastic errors and general H

When the errors on the estimate for f are bounded and deterministic, we gain no advantage by increasing the length of the estimation horizon H. However, when the errors are centered and stochastic, the estimation horizon now plays a critical roll. Increasing H allows for concentration of the errors, leading to overall better estimates for f. At the same time, increasing H causes the deterministic bias from our finite difference approximations to increase. In the following section, we show how to balance these two factors and choose an optimal H. First, we state our main theorem.

Theorem 13. *With step size*

$$\eta = \frac{g^{2/3}}{M^{1/2}G^{5/3}\tau^{1/3}(\log\frac{T}{\gamma})^{1/6}T^{5/6}}$$

and estimation horizon

$$H = \frac{\tau^{2/5}(\log\frac{T}{\gamma})^{1/5}}{M^{2/5}g^{4/5}}\eta^{-4/5}$$

the iterates of PerfGD satisfy

$$\min_{1\le t\le T}|\nabla\mathcal{L}_t|^2 = \max\left\{\mathcal{O}\left(\ell_{\max}\left[\frac{\tau^{1/3}}{g^{2/3}T^{1/6}}\cdot M^{1/2}G^{5/3}(\log\frac{T}{\gamma})^{1/6}\right]\right), \mathcal{O}\left(g^2 + \ell_{\max}\cdot\text{poly}(M, G, \log\frac{T}{\gamma})\cdot\frac{\tau^{3/5}}{g^{2/3}T^{1/6}}\right)\right\}$$

with probability at least $1 - \mathcal{O}(\gamma)$ *as* $\tau \to 0$ *and* $T \to \infty$.

We remark briefly that we choose to analyze $\tau \to 0$ since if the estimates for f_t are computed from random samples of increasing size, then we expect the variance of these estimates (measured by τ) to decay to zero as the sample size $n \to \infty$. For instance, for estimating the mean of a Gaussian we will have $\tau^2 = \mathcal{O}(1/n)$.

The proof of Theorem 13 follows from two key lemmas.

Lemma 14. *If* X *is* τ^2*-subgaussian and* Y *is any random variable with* $|Y| \le B$ *w.p. 1, then* XY *is* $B^2\tau^2$*-subgaussian.*

A critical fact about this lemma is that the random variables involved need not be independent.

Proof. By definition, Z is s^2-subgaussian if $\mathbb{E}e^{Z^2/s^2} \leq 2$. Observe that since the exponential function is monotonic, we have

$$\mathbb{E}e^{X^2Y^2/B^2\tau^2} \leq \mathbb{E}e^{X^2B^2/B^2\tau^2} = \mathbb{E}e^{X^2/\tau^2} \leq 2.$$

Thus XY is $B^2\tau^2$-subgaussian.

□

Lemma 15. *We have $\widehat{f'_t} = f'_t + b_t + e_t$, where b_t is a deterministic bias term with $|b_t| = \mathcal{O}(MGH\eta)$. Under the additional assumption that θ_t converge monotonically, e_t is $\mathcal{O}\left(\frac{G^2\tau^2}{H^3g^4\eta^2}\right)$-subgaussian.*

Proof. The pseudoinverse used to compute $\widehat{f'_t}$ is equivalent to solving the least-squares problem

$$\widehat{f'_t} = \underset{\alpha}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{H}(\alpha(\theta_{t-i}-\theta_t) - (\hat{f}_{t-i}-\hat{f}_t))^2 \implies \widehat{f'_t} = \frac{\sum_{i=1}^{H}(\hat{f}_{t-i}-\hat{f}_t)(\theta_{t-i}-\theta_t)}{\sum_{i=1}^{H}(\theta_{t-i}-\theta_t)^2}. \tag{36}$$

Writing $\hat{f}_t = f_t + \varepsilon_t$ with ε_t τ-subgaussian, we can apply Taylor's theorem to rewrite

$$\hat{f}_{t-i} - \hat{f}_t = f'_t(\theta_{t-i}-\theta_t) + \frac{1}{2}f''(\xi_i)(\theta_{t-i}-\theta_t)^2 + \varepsilon_{t-i} - \varepsilon_t. \tag{37}$$

Using the explicit solution in (36) and substituting (37) for $\hat{f}_{t-i} - \hat{f}_t$, we find that

$$|\widehat{f'_t} - f'_t| \leq \underbrace{\frac{\frac{1}{2}\sum_{i=1}^{H}|f''(\xi_i)||\theta_{t-i}-\theta_t|^3}{\sum_{i=1}^{H}(\theta_{t-i}-\theta_t)^2}}_{b_t} + \underbrace{\frac{\sum_{i=1}^{H}(\varepsilon_{t-i}-\varepsilon_t)(\theta_{t-i}-\theta_t)}{\sum_{i=1}^{H}(\theta_{t-i}-\theta_t)^2}}_{e_t}.$$

To bound b_t, observe that since $\theta_t = \theta_{t-i} - \eta(\hat{\nabla}\mathcal{L}_{t-i} + \cdots + \hat{\nabla}\mathcal{L}_{t-1})$ and $|\hat{\nabla}\mathcal{L}_s| \leq G$ and $i \leq H$, we have $|\theta_{t-i} - \theta_t| \leq HG\eta$ for all i, t. Since $|f''(\xi_i)| \leq M$, we have

$$\begin{aligned} |b_t| &\leq \frac{\frac{1}{2}\sum_{i=1}^{H}MHG\eta(\theta_{t-i}-\theta_t)^2}{\sum_{i=1}^{H}(\theta_{t-i}-\theta_t)^2} \\ &= \mathcal{O}(MGH\eta). \end{aligned}$$

Next we bound e_t. Since we have assumed that θ_t converge monotonically and $|\hat{\nabla}\mathcal{L}_t| \geq g$, we have

$$\frac{1}{\sum_{i=1}^{H}(\theta_{t-i}-\theta_t)^2} \leq \frac{1}{\sum_{i=1}^{H}(ig\eta)^2} = \mathcal{O}(\frac{1}{H^3g^2\eta^2}).$$

In the numerator, we have

$$\left|\sum_{i=1}^{H}(\varepsilon_{t-i}-\varepsilon_t)(\theta_{t-i}-\theta_t)\right| \leq HG\eta\sum_{i=1}^{H}|\varepsilon_{t-i}-\varepsilon_t|.$$

Combining these, we have

$$e_t = \mathcal{O}\left(\frac{G}{H^2g^2\eta}\right)\sum_{i=1}^{H}|\varepsilon_{t-i}-\varepsilon_t|. \tag{38}$$

We make the additional simplifying assumption that the $|\varepsilon_{t-i} - \varepsilon_t|$ are independent. We can accomplish this splitting our dataset drawn from $\mathcal{D}(\theta_t)$ into H parts and estimating f_t once with each component, then replacing the terms $(\hat{f}_{t-i} - \hat{f}_t)$ with $(\hat{f}_{t-i} - \hat{f}_{t,i})$ in equation (36), where $\hat{f}_{t,i}$ is the estimate of f_t from the i-th partition of the dataset. The errors ε_t in the expression for e_t now become independent copies $\varepsilon_{t,i}$, and the terms in equation (38) are indeed independent.

Under this assumption, $|\varepsilon_{t-i} - \varepsilon_{t,i}|$ are independent $2\tau^2$-subgaussian random variables. Their sum is therefore $\sum_{i=1}^{H} 2\tau^2 = \mathcal{O}(H\tau^2)$-subgaussian. Finally, by Lemma 14, it follows that e_t is $\mathcal{O}(\frac{G^2\tau^2}{H^3 g^4 \eta^2})$-subgaussian.

□

With these two lemmas, we can now prove the main theorem. The structure of the proof is similar to that of Theorem 2.

Proof of Theorem 13. We first establish a high-probability bound on $|e_t|$. By the subgaussian tail bound and a union bound over $t = 1$ to T, a simple calculation shows that

$$|e_t| = \mathcal{O}\left(\frac{G\tau\sqrt{\log\frac{T}{\gamma}}}{g^2\eta H^{3/2}}\right)$$

with probability at least $1-\gamma$ for all $t = 1, \ldots, T$. Combining this bound with the bound on $|b_t|$ from Lemma 15, we find that

$$|\widehat{f}_t' - f_t'| = \mathcal{O}\left(MG\eta H + \frac{G\tau\sqrt{\log\frac{T}{\gamma}}}{g^2\eta}H^{-3/2}\right).$$

With H chosen as is in the theorem, this bound simplifies to

$$|\widehat{f}_t' - f_t'| = \mathcal{O}\left(\frac{M^{3/5}G\tau^{2/5}(\log\frac{T}{\gamma})^{1/5}}{g^{4/5}}\eta^{1/5}\right) \equiv \mathbf{E}_1. \tag{39}$$

From the proof of Lemma 1, we know that

$$|\hat{\nabla}\mathcal{L}_t - \nabla\mathcal{L}_t| = \mathcal{O}\left(\ell_{\max}\left[\mathbf{E}_1 + F\delta\sqrt{\log\frac{1}{\delta}}\right]\right), \tag{40}$$

where δ is a (high-probability) bound on the error of f_t. Again assuming that this error is τ^2-subgaussian, we have that

$$(\text{error on } f_t) = \mathcal{O}\left(\tau\sqrt{\log\frac{T}{\gamma}}\right)$$

for all $t = 1, \ldots, T$ with probability at least $1-\gamma$. Thus we can take $\delta = \tau\sqrt{\log(T/\gamma)}$, in which case the second term in equation (40) is $\mathcal{O}(\mathbf{E}_1)$ as $\tau \downarrow 0$. It follows that $|\hat{\nabla}\mathcal{L}_t - \nabla\mathcal{L}_t| = \mathcal{O}(\ell_{\max}\mathbf{E}_1) \equiv \mathbf{E}_2$ with high probability.

Finally, by the same analysis used in the proof of Theorem 2, we have that

$$\min_{1\leq t\leq T} |\nabla\mathcal{L}_t|^2 = \mathcal{O}\left(\frac{\ell_{\max} + \eta GT\mathbf{E}_2}{T\eta}\right).$$

Choosing η as in the theorem statement and substituting our bound on $\mathbf{E}_2$ yields the desired result. The max in the theorem statement follows from the same logic as in Theorem 2 plus the bound on the error performative gradient error $\mathbf{E}_2$. We also note that similarly to Theorem 2, we can choose the stopping criterion to match the leading order behavior in the two terms in the max in the theorem statement so that $\min_{1\leq t\leq T} |\nabla\mathcal{L}_t|^2 \to 0$ as $\tau \to 0$ and $T \to \infty$. □

F. Experiment details

In all of the following experiments, whenever the stated estimation horizon H is longer than the entire history on a particular iteration of PerfGD, we simply use $H =$ length of the existing history for that iteration instead. Furthermore, in all of the experiments, RGD, FLX, and PerfGD were all run using a learning rate of $\eta = 0.1$. Code for reproducing these experiments can be found at `https://github.com/zleizzo/PerfGD`.

F.1. Mixture of Gaussians and nonlinear mean (§5.1)

For the nonlinear mean experiment, we set $a_0 = a_1 = 1$ and $\sigma^2 = 1$. At each iteration, we drew $n = 500$ data points. We initialized PerfGD using only one step of RGD, and at each step after the initialization we used the previous $H = 4$ steps to estimate $f'(\theta)$. For FLX, we tried $\delta \in \{0.1, 0.3, 1\}$ and selected the best result, which was $\delta = 0.3$. The analytical values for θ_{OPT} and θ_{STAB} are given by

$$\theta_{\mathrm{OPT}} = -\frac{2a_0}{3a_1}, \qquad \theta_{\mathrm{STAB}} = -\frac{a_0}{a_1}.$$

For the Gaussian mixture experiment, we set $\gamma = 0.5$, $\sigma_1^2 = 1$, $a_{1,0} = -0.5$, $a_{1,1} = 1$, $s_2^2 = 0.25$, $a_{2,0} = 1$, and $a_{2,1} = -0.3$. At each iteration, we drew $n = 1000$ data points. We initialized PerfGD using only one step of RGD, and at each step after the initialization we use the entire history to estimate $f_i'(\theta)$. For FLX, we tried $\delta \in \{0.1, 0.3, 1\}$ and selected the best result, which was $\delta = 0.1$. The analytical values for θ_{OPT} and θ_{STAB} are given by

$$\theta_{\mathrm{OPT}} = -\frac{1}{2}\frac{\gamma a_{1,0} + (1-\gamma)a_{2,0}}{\gamma a_{1,1} + (1-\gamma)a_{2,1}}, \qquad \theta_{\mathrm{STAB}} = \frac{\gamma a_{1,0} + (1-\gamma)a_{2,0}}{\gamma a_{1,1} + (1-\gamma)a_{2,1}}.$$

F.2. Pricing (§5.2)

We set $d = 5$ for this experiment. We then set $\mu_0 = 6 \cdot \mathbf{1} + \mathrm{Unif}[0,1]^5$ with a fixed random seed; in this case, it came out to $\mu_0 \approx [6.55, 6.72, 6.60, 6.54, 6.42]^\top$. We set $\Sigma = I \in \mathbb{R}^{5\times 5}$ (i.e. the 5×5 identity matrix) and $\varepsilon = 1.5$. At each iteration, we drew $n = 500$ data points. We initialized PerfGD with 14 steps of RGD, and at each step after initialization we used the previous $H = 25$ steps to estimate $df/d\theta$. For FLX, we first searched over $\delta \in \{3, 4, 5, 6\}$, then refined the grid and searched over $\delta \in \{4.25, 4.5, 4.75\}$. We selected the best value $\delta = 4.75$. The analytical values for θ_{OPT} and θ_{STAB} are given by

$$\theta_{\mathrm{OPT}} = \frac{\mu_0}{2\varepsilon}, \qquad \theta_{\mathrm{STAB}} = \frac{\mu_0}{\varepsilon}.$$

We also plot the minimum singular value of $\Delta\theta$ over the course of the trajectory. It remains bounded away from 0, so the lower bound assumption on the minimum singular value in Theorem 5 is valid in this case.

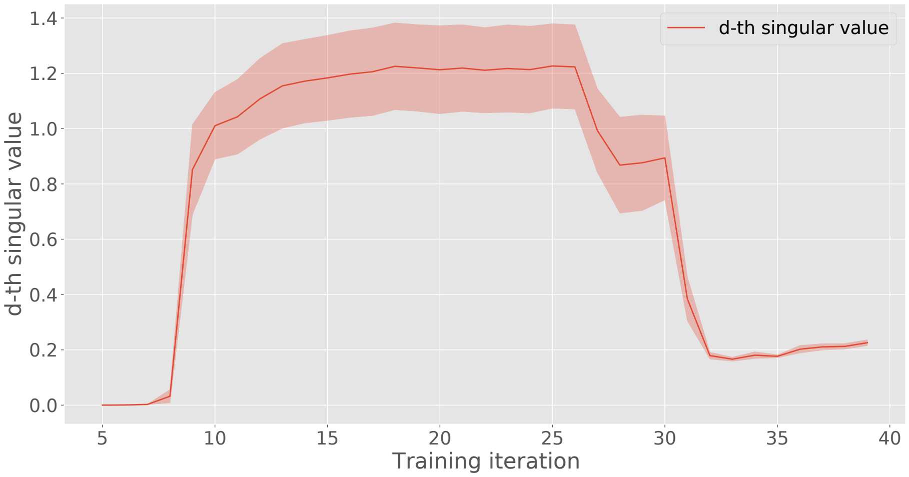

Figure 6. Minimum singular value of $\Delta\theta$ over the course of training.

F.3. Binary classification (§5.3)

Here the features $x \in \mathbb{R}$ are one-dimensional, while our model parameters $\theta \in \mathbb{R}^2$ allow for a bias term. We set $\sigma_0^2 = 0.25$, $\mu_0 = 1$, $\sigma_1^2 = 0.25$, $\mu_1 = -1$, and $\varepsilon = 3$. The regularization strength for ℓ was $\lambda = 10^{-2}$, i.e.

$$\ell(x, y; \theta) = -y \log h_\theta(x) - (1-y)\log(1 - h_\theta(x)) + \frac{10^{-2}}{2}\|\theta\|^2.$$

When approximating the derivatives of the means of the mixtures with respect to θ, we assume that it is known that the derivative of the non-spam email mean is independent of θ, and we also assume knowledge of the fact that the mean of the spam email features depends only on θ_1 (i.e. the non-bias parameter). At each iteration, we drew $n = 500$ data points. We initialize PerfGD using only one step of RGD, and at each step after the initialization we use the entire history to estimate $f'(\theta)$. For FLX, despite an extensive grid search over $\delta \in \{0.1, 0.2, 0.3, \dots, 4.0\}$, FLX was unable to converge.

F.4. Regression (§5.4)

We set $\mu_x = 1.67$, $\sigma_x^2 = 1$, $a_0 = a_1 = 1.67$, and regularization strength $\lambda = 3.33$ for the loss, i.e.

$$\ell(x, y; \theta) = \frac{1}{2}(\theta x - y)^2 + \frac{3.33}{2}|\theta|^2.$$

The variance of $y|x$ was set to 4.12. At each iteration, we drew $n = 500$ data points and used the entire history to estimate $d\beta/d\theta$. For FLX, we searched over $\delta \in \{1.5, 2, 2.5, 3, 3.5, 4\}$, then selected the best value $\delta = 2$.The analytical values for θ_{OPT} and θ_{STAB} are given by

$$\theta_{\text{OPT}} = \frac{c \cdot a_0}{c \cdot (1 - a_1) + \frac{\lambda}{1 - a_1}}, \qquad \theta_{\text{STAB}} = \frac{c \cdot a_0}{c \cdot (1 - a_1) + \lambda},$$

where $c = \mu_x^2 + \sigma_x^2$.