Supplement to "The Implicit Regularization of Stochastic Gradient Flow for Least Squares"

Alnur Ali Edgar Dobriban Ryan J. Tibshirani

This supplementary document contains additional details, proofs, and experiments for the paper "The Implicit Regularization of Stochastic Gradient Flow for Least Squares". All section, figure, and equation numbers in this document begin with the letter "S", to differentiate them from those appearing in the main paper (which appear without the prepended letter "S").

Proof of Lemma 1

For simplicity, below we will omit the source of the randomness for the various estimators. Implicitly, the randomness is from minibatching in SGD, and from the normal random increments in the discretization of SGD (which we wil can dSGF).

By taking expectations in the SGD iteration, we find

$$\mathbb{E}\beta^{(k)} = \mathbb{E}\beta^{(k-1)} + \epsilon \cdot \mathbb{E}\frac{1}{n}X^T(y - X\beta^{(k-1)}).$$

This identity only uses that the stochastic gradients are unbiased estimators for the true gradients. Thus, it is true even more generally for any loss function, not just for quadratic loss. However, for quadratic loss, we have a very special property, namely that the *gradient is linear* in the parameter. Using this, we can move the expectation inside, and we find

$$\mathbb{E}\beta^{(k)} = \mathbb{E}\beta^{(k-1)} + \epsilon \cdot \frac{1}{n}X^T(y - X\mathbb{E}\beta^{(k-1)}).$$

From this, it follows by a direct induction argument that $\mathbb{E}\beta^{(k)} = \beta_{gd}^{(k)}$, where $\beta_{gd}^{(k)}$ is the gradient descent iteration with learning rate ϵ started from 0. Indeed, for $k = 0$, we have $\mathbb{E}\beta^{(k)} = \beta_{gd}^{(k)} = 0$. Next, the two sequences satisfy the same recurrence. Hence the induction finishes the argument.

A similar reasoning holds for dSGF. This shows that $\mathbb{E}_{\tilde{Z}}\tilde{\beta}^{(k)} = \mathbb{E}_{\mathcal{I}_1,\ldots,\mathcal{I}_k}\beta^{(k)}$.

By taking the covariance of the SGD iteration, conditionally on the previous iterate $\beta^{(k-1)}$, we find

$$\begin{aligned}\mathrm{Cov}[\beta^{(k)}|\beta^{(k-1)}] &= \epsilon^2 \cdot \mathrm{Cov}\left[\frac{1}{m}X_{\mathcal{I}_k}^T(y_{\mathcal{I}_k} - X_{\mathcal{I}_k}\beta^{(k-1)}) - \frac{1}{n}X^T(y - X\beta^{(k-1)})\right] \\ &= \epsilon \cdot Q_\epsilon(\beta^{(k-1)}).\end{aligned}$$

Note that the definition of Q already includes an ϵ. This shows that the conditional covariance of the SGD iterate, conditoned on the previous iterate, is the same as for dSGF. As before, this observation holds not just for quadratic objectives, but also for general objectives. However, noting that Q is a *quadratic function* of the parameter, it follows from an inductive argument that the covariance matrix of the iterates $\beta^{(k)}$ and $\tilde{\beta}^{(k)}$ equals at every iteration. Indeed, the reason is that the covariance at each iteration only depends on second order statistics of the previous iteration (including the mean and the covariance), and so the induction step will hold. This shows that $\mathrm{Cov}_{\tilde{Z}}\tilde{\beta}^{(k)} = \mathrm{Cov}_{\mathcal{I}_1,\ldots,\mathcal{I}_k}\beta^{(k)}$, finishing the proof.

We can also find the explicit form of the recursion. While this is not required in the statement of the lemma, it is used in our numerical examples.

$$\begin{aligned}\mathrm{Cov}[\beta^{(k)}] &= \mathbb{E}\mathrm{Cov}[\beta^{(k)}|\beta^{(k-1)}] \\ &= \frac{\epsilon^2}{mn}\sum_{i=1}^{n}\mathbb{E}(y_i - x_i^T\beta^{(k-1)})^2 x_i x_i^T - \frac{\epsilon^2}{mn^2}\mathbb{E}(\sum_{i=1}^{n}\mathbb{E}(y_i - x_i^T\beta^{(k-1)})x_i)^{\otimes 2}\end{aligned}$$

For the first term, we can write

$$\begin{aligned}&= \frac{\epsilon^2}{mn}\sum_{i=1}^{n}\mathbb{E}(y_i - x_i^T\mathbb{E}\beta^{(k-1)} + x_i^T\mathbb{E}\beta^{(k-1)} - x_i^T\beta^{(k-1)})^2 x_i x_i^T \\ &= \frac{\epsilon^2}{mn}\sum_{i=1}^{n}(y_i - x_i^T\beta_{gd}^{(k-1)})^2 x_i x_i^T + \frac{\epsilon^2}{mn}\sum_{i=1}^{n}[x_i^T(\mathbb{E}\beta^{(k-1)} - \beta^{(k-1)})]^2 x_i x_i^T \\ &= \frac{\epsilon^2}{mn}\sum_{i=1}^{n}(y_i - x_i^T\beta_{gd}^{(k-1)})^2 x_i x_i^T + \frac{\epsilon^2}{mn}\sum_{i=1}^{n}x_i^T\mathrm{Cov}[\beta^{(k-1)}]x_i \cdot x_i x_i^T.\end{aligned}$$

For the first term, we can write

$$\begin{aligned}\sum_{i=1}^{n}\mathbb{E}(y_i - x_i^T\beta^{(k-1)})x_i &= \sum_{i=1}^{n}(y_i - x_i^T\mathbb{E}\beta^{(k-1)})x_i \\ &= \sum_{i=1}^{n}(y_i - x_i^T\beta_{gd}^{(k-1)})x_i + \sum_{i=1}^{n}x_i^T[\beta_{gd}^{(k-1)} - \mathbb{E}\beta^{(k-1)}]x_i \\ &= \sum_{i=1}^{n}(y_i - x_i^T\beta_{gd}^{(k-1)})x_i + \sum_{i=1}^{n}n\hat{\Sigma}[\beta_{gd}^{(k-1)} - \mathbb{E}\beta^{(k-1)}]\end{aligned}$$

so

$$\begin{aligned}&\frac{\epsilon^2}{mn^2}\mathbb{E}(\sum_{i=1}^{n}\mathbb{E}(y_i - x_i^T\beta^{(k-1)})x_i)^{\otimes 2} \\ &= \frac{\epsilon^2}{mn^2}[\sum_{i=1}^{n}(y_i - x_i^T\beta_{gd}^{(k-1)})x_i]^{\otimes 2} + \frac{\epsilon^2}{m}\mathbb{E}[\hat{\Sigma}[\beta_{gd}^{(k-1)} - \mathbb{E}\beta^{(k-1)}]]^{\otimes 2} \\ &= \frac{\epsilon^2}{mn^2}[\sum_{i=1}^{n}(y_i - x_i^T\beta_{gd}^{(k-1)})x_i]^{\otimes 2} + \frac{\epsilon^2}{m}\hat{\Sigma}\mathrm{Cov}[\beta^{(k-1)}]\hat{\Sigma}.\end{aligned}$$

This gives an explicit linear recursion for the covariance matrices. The first term can be viewed as a covariance matrix of the gradients evaluated at the mean value of the process (i.e., at the value of the GD iteration). The second term depends on the covariance of the previous iteration.

Proof of Lemma 2

As the diffusion coefficient $Q_\epsilon(\beta(t))^{1/2}$ is Lipschitz continuous and positive semidefinite, standard results from numerical analysis (e.g., Øksendal (2003)) show that the solution to the differential equation (4) exists and is unique.

Now consider the process $\tilde{\beta}(t) = \exp(t\hat{\Sigma})\beta(t)$. By Ito's lemma,

$$d\tilde{\beta}(t) = \hat{\Sigma}\exp(t\hat{\Sigma})\beta(t)dt + \exp(t\hat{\Sigma})d\beta(t).$$

Plugging in the expression for $d\beta(t)$ from (4) and simplifying, we see that

$$d\tilde{\beta}(t) = \exp(t\hat{\Sigma})\Big(\frac{1}{n}X^T y\Big)dt + \exp(t\hat{\Sigma})Q_\epsilon(\beta(t))^{1/2}dW(t),$$

or, equivalently,

$$\tilde{\beta}(t) = \int_0^t \exp(\tau\hat{\Sigma})\Big(\frac{1}{n}X^T y\Big)d\tau + \int_0^t \exp\Big(\tau\hat{\Sigma}\Big)Q_\epsilon(\beta(\tau))^{1/2}dW(\tau).$$

Changing variables back yields

$$\beta(t) = \exp(-t\hat{\Sigma}) \cdot \int_0^t \exp(\tau\hat{\Sigma})\Big(\frac{1}{n}X^T y\Big)d\tau + \exp(-t\hat{\Sigma}) \cdot \int_0^t \exp(\tau\hat{\Sigma})Q_\epsilon(\beta(\tau))^{1/2}dW(\tau).$$

Considering only the first integral above, by arguments similar to those given in Lemma 1 of Ali et al. (2018), we obtain

$$\int_0^t \exp(\tau\hat{\Sigma})\Big(\frac{1}{n}X^T y\Big)d\tau = \big(\exp(t\hat{\Sigma}) - I\big)(X^T X)^+ X^T y,$$

and so

$$\exp(-t\hat{\Sigma}) \cdot \int_0^t \exp(\tau\hat{\Sigma})\Big(\frac{1}{n}X^T y\Big)d\tau = (X^T X)^+\big(I - \exp(-t\hat{\Sigma})\big)X^T y = \hat{\beta}^{\mathrm{gf}}(t),$$

which gives the result. □

Proof of Lemma 3

To keep things simple, we prove the result using the uncentered covariance matrix of the stochastic gradients, i.e., we let $Q_\epsilon(\hat{\beta}^{\mathrm{sgf}}(t)) = (1/(nm))X^T F(\hat{\beta}^{\mathrm{sgf}}(t))X$, where $h(\beta) = (y_1 - x_1^T\beta, \ldots, y_n - x_n^T\beta)$ are the residuals at β, and $F(\beta) = \mathrm{diag}(h(\beta))^2$. A similar result holds for the actual covariance matrix, but it is a little difficult to interpret.

Calculations similar to those given in Lemma 4 (appearing below) show

$$\mathbb{E}_Z\|\hat{\beta}^{\mathrm{sgf}}(t) - \tilde{\beta}^{\mathrm{sgf}}(t)\|_2^2 = \mathbb{E}\int_0^t \mathrm{tr}\left[\exp((\tau-t)\hat{\Sigma})\left(Q_\epsilon(\hat{\beta}^{\mathrm{sgf}}(\tau))^{1/2} - \Big(\frac{\epsilon}{m}\cdot\hat{\Sigma}\Big)^{1/2}\right)^2\exp((\tau-t)\hat{\Sigma})\right]d\tau.$$

Continuing on, and writing $L = \lambda_{\max}(\hat{\Sigma})$, we have

$$\begin{aligned}
\mathbb{E}_Z\|\hat{\beta}^{\mathrm{sgf}}(t) - \tilde{\beta}^{\mathrm{sgf}}(t)\|_2^2 &= \mathbb{E}_Z\int_0^t \mathrm{tr}\left[\left(Q_\epsilon(\hat{\beta}^{\mathrm{sgf}}(\tau))^{1/2} - \Big(\frac{\epsilon}{m}\cdot\hat{\Sigma}\Big)^{1/2}\right)^2\exp(2(\tau-t)\hat{\Sigma})\right]d\tau \\
&= \mathbb{E}_Z\int_0^t \mathrm{tr}\left[V^T\left(Q_\epsilon(\hat{\beta}^{\mathrm{sgf}}(\tau))^{1/2} - \Big(\frac{\epsilon}{m}\cdot\hat{\Sigma}\Big)^{1/2}\right)^2 V\exp(2(\tau-t)S)\right]d\tau \\
&\leq \mathbb{E}_Z\int_0^t \mathrm{tr}\left[V^T\left(Q_\epsilon(\hat{\beta}^{\mathrm{sgf}}(\tau))^{1/2} - \Big(\frac{\epsilon}{m}\cdot\hat{\Sigma}\Big)^{1/2}\right)^2 V\right]\mathrm{tr}\left[\exp(2(\tau-t)S)\right]d\tau \\
&\leq \frac{4Lp^2\epsilon}{m}\cdot\int_0^t\sum_{i=1}^n \mathbb{E}_Z\Big[\big|(y_i - x_i^T\hat{\beta}^{\mathrm{sgf}}(\tau))^2 - 1\big|\Big]\mathrm{tr}\left[\exp(2(\tau-t)\hat{\Sigma})\right]d\tau \\
&\leq \frac{4Lp^3\epsilon}{m}\cdot\int_0^t\sum_{i=1}^n \mathbb{E}_Z\Big[\big|(y_i - x_i^T\hat{\beta}^{\mathrm{sgf}}(\tau))^2 - 1\big|\Big]d\tau,
\end{aligned}$$

where we used the eigendecomposition $\hat{\Sigma} = VSV^T$ on the second line, the helper Lemma S.1 (appearing below) on the third, the helper Lemma S.2 (appearing below) on the fourth, and the fact that the map $A \mapsto \mathrm{tr}\,\exp(A)$ is operator monotone on the fifth. This proves the result. □

Lemma S.1. *Let $A \in \mathbb{R}^{n\times n}$ be a nonnegative diagonal matrix, and $B \in \mathbb{R}^{n\times n}$ be a positive semidefinite matrix. Then* $\mathrm{tr}\,(AB) \leq \mathrm{tr}\,(A)\mathrm{tr}\,(B)$.

Proof. Write $\operatorname{tr}(AB) = \sum_{i=1}^n A_{ii}B_{ii}$. Cauchy-Schwarz shows that

$$\sum_{i=1}^n A_{ii}B_{ii} \le \Big(\sum_{i=1}^n A_{ii}^2\Big)^{1/2}\Big(\sum_{i=1}^n B_{ii}^2\Big)^{1/2}.$$

Using the simple fact that $\|x\|_2 \le \|x\|_1$, along with the fact that A, B have nonnegative diagonal entries, now yields the result. □

Lemma S.2. *Fix y, X and β. Let $h(\beta) = (y_1 - x_1^T\beta, \dots, y_n - x_n^T\beta)$ denote the residuals at β, and $F(\beta) = \operatorname{diag}(h(\beta))^2$. Then,*

$$\operatorname{tr}\left[\left(Q_\epsilon(\beta)^{1/2} - \Big(\frac{\epsilon}{m}\cdot\hat\Sigma\Big)^{1/2}\right)^2\right] \le \frac{4Lp^2\epsilon}{m}\cdot\operatorname{tr}\big[|F(\beta) - I|\big],$$

where the absolute value is to be interpreted elementwise.

Proof. Using the matrix perturbation inequality given in Lemma A.2 of Nguyen et al. (2019), we see that

$$\begin{aligned}\operatorname{tr}\left[\left(Q_\epsilon(\beta)^{1/2} - \Big(\frac{\epsilon}{m}\cdot\hat\Sigma\Big)^{1/2}\right)^2\right] &= \left\|Q_\epsilon(\beta)^{1/2} - \Big(\frac{\epsilon}{m}\cdot\hat\Sigma\Big)^{1/2}\right\|_F^2 \\ &\le p\cdot\left\|Q_\epsilon(\beta)^{1/2} - \Big(\frac{\epsilon}{m}\cdot\hat\Sigma\Big)^{1/2}\right\|_2^2 \\ &\le 4p^2\cdot\left\|Q_\epsilon(\beta) - \frac{\epsilon}{m}\cdot\hat\Sigma\right\|_2. \end{aligned} \tag{S.1}$$

Noting the expression for the covariance matrix of the stochastic gradients given in (S.3), we obtain for (S.1) that

$$\begin{aligned}4p^2\cdot\left\|Q_\epsilon(\beta) - \frac{\epsilon}{m}\cdot\hat\Sigma\right\|_2 &= \frac{4p^2\epsilon}{m}\cdot\left\|\frac{1}{n}X^T\big(F(\beta) - I\big)X\right\|_2 \\ &\le \frac{4Lp^2\epsilon}{m}\cdot\|F(\beta) - I\|_2 \\ &\le \frac{4Lp^2\epsilon}{m}\cdot\operatorname{tr}\big[|F(\beta) - I|\big].\end{aligned}$$

Here, we let $L = \lambda_{\max}(\hat\Sigma)$, $h(\beta) = (y_1 - x_1^T\beta, \dots, y_n - x_n^T\beta)$ denote the residuals at β, and $F(\beta) = \operatorname{diag}(h(\beta))^2$. This shows the result. □

Proof of Lemma 4

As $\hat\beta^{\mathrm{gf}}(t)$ is constant and the Brownian motion term in (4) has mean zero, we have

$$\begin{aligned}\operatorname{tr}\operatorname{Cov}_Z(\hat\beta^{\mathrm{sgf}}(t)) = \operatorname{tr}\mathbb{E}_Z\Bigg[\exp(-t\hat\Sigma)\Bigg(\int_0^t \exp(\tau\hat\Sigma)Q_\epsilon(\hat\beta^{\mathrm{sgf}}(\tau))^{1/2}dW(\tau)\Bigg) \\ \Bigg(\int_0^t \exp(\tau\hat\Sigma)Q_\epsilon(\hat\beta^{\mathrm{sgf}}(\tau))^{1/2}dW(\tau)\Bigg)^T\exp(-t\hat\Sigma)\Bigg].\end{aligned}$$

Using Ito's isometry along with the linearity of the trace, we obtain

$$\operatorname{tr}\operatorname{Cov}_Z(\hat\beta^{\mathrm{sgf}}(t)) = \mathbb{E}_Z\int_0^t \operatorname{tr}\Big[\exp((\tau - t)\hat\Sigma)Q_\epsilon(\hat\beta^{\mathrm{sgf}}(\tau))\exp((\tau - t)\hat\Sigma)\Big]d\tau. \tag{S.2}$$

For the squared error loss, the covariance matrix of the stochastic gradients at β sampled with replacement has a relatively well-known simplified form (cf. Hoffer et al. (2017); Zhang et al. (2017); Hu et al. (2017)). Let $h(\beta) = (y_1 - x_1^T\beta, \ldots, y_n - x_n^T\beta)$ denote the residuals at β, $F(\beta) = \mathrm{diag}(h(\beta))^2$, and $\tilde{F}(\beta) = n^{-1}h(\beta)h(\beta)^T$. Then,

$$\begin{aligned} Q_\epsilon(\beta) &= \mathrm{Cov}_{\mathcal{I}}\left(\frac{1}{m}X_{\mathcal{I}}^T(y_{\mathcal{I}} - X_{\mathcal{I}}\beta)\right) \\ &= \frac{1}{nm}X^T(F(\beta) - \tilde{F}(\beta))X \\ &\preceq \frac{1}{nm}X^T F(\beta)X. \end{aligned} \tag{S.3}$$

Letting $A = \exp((\tau - t)\hat{\Sigma})$, the trace appearing in (S.2) may be expressed as

$$\frac{\epsilon}{mn} \cdot \mathrm{tr}\left(AX^T F(\hat{\beta}^{\mathrm{sgf}}(\tau))XA\right) = \frac{\epsilon}{mn} \cdot \mathrm{tr}\left(F(\hat{\beta}^{\mathrm{sgf}}(\tau))XA^2X^T\right).$$

Since XA^2X^T is positive semidefinite, the matrix $F(\hat{\beta}^{\mathrm{sgf}}(\tau))XA^2X^T$ is the product of a nonnegative diagonal matrix and a positive semidefinite matrix; this satisfies the conditions of Lemma S.1, which yields the bound

$$\mathrm{tr}\left(F(\hat{\beta}^{\mathrm{sgf}}(\tau))XA^2X^T\right) \leq \mathrm{tr}\left(F(\hat{\beta}^{\mathrm{sgf}}(\tau))\right)\mathrm{tr}\left(XA^2X^T\right).$$

By straightforward manipulations, we see $(1/n)\mathrm{tr}\left(XA^2X^T\right) = \mathrm{tr}\left(\hat{\Sigma}\exp(2(\tau - t)\hat{\Sigma})\right)$. Therefore, $\mathrm{tr}\,\mathrm{Cov}_Z(\hat{\beta}^{\mathrm{sgf}}(t))$, as in (S.2), may be bounded as

$$\begin{aligned} \mathrm{tr}\,\mathrm{Cov}_Z(\hat{\beta}^{\mathrm{sgf}}(t)) &\leq \frac{\epsilon}{m} \cdot \mathbb{E}_Z \int_0^t \mathrm{tr}\left(F(\hat{\beta}^{\mathrm{sgf}}(\tau))\right)\mathrm{tr}\left(\hat{\Sigma}\exp(2(\tau - t)\hat{\Sigma})\right)d\tau \\ &= \frac{\epsilon}{m} \cdot \int_0^t \mathbb{E}_Z\left[\mathrm{tr}\left(F(\hat{\beta}^{\mathrm{sgf}}(\tau))\right)\right]\mathrm{tr}\left(\hat{\Sigma}\exp(2(\tau - t)\hat{\Sigma})\right)d\tau. \end{aligned}$$

The equality followed by Fubini's theorem (which applies here, since the product of the trace of a nonnegative diagonal matrix and the trace of a positive semidefinite matrix, is nonnegative). As $\mathbb{E}_Z\left[\mathrm{tr}\left(F(\hat{\beta}^{\mathrm{sgf}}(\tau))\right)\right] = 2nf(\hat{\beta}^{\mathrm{sgf}}(\tau))$, this shows the result. □

Proof of Lemma 5

In this lemma, it will be helpful to start slightly more generally, with the SDE for SGF on a general loss function g. The specific proofs of this lemma are in Sections S.0.1 and S.0.2.

To approximate discrete time SGD with learning rate ϵ and batch size m, it is not hard to see that the same logic we have used throughout the paper leads to the SDE

$$d\beta(t) = -\nabla g(\beta(t))dt + \eta\sigma(\beta(t))dW(t)$$

where $\sigma(\beta(t))\sigma(\beta(t))^T$ is the covariance of the gradients at parameter value $\beta(t)$, and $\eta = \epsilon/\sqrt{m}$.

We derive the SDE for the behavior of the loss function itself, for a general loss. For gradient flow on a loss function g, i.e., the dynamics $d\beta(t) = -\nabla g(\beta(t))dt$, it is well known that the dynamics induced on the loss function is:

$$dg(\beta(t)) = -|\nabla g(\beta(t))|^2 dt.$$

This shows that the loss function is always non-increasing, i.e., that gradient flow is a descent method. In contrast, we will find that the loss for stochastic gradient flow is *not* always non-increasing. We mention that a related calculation has been performed in (Zhu et al., 2018), under different assumptions (starting from a local min, integrating over time), and for a different purpose (to understand dynamics escaping local minima).

Proposition 1 (Dynamics of the loss for SGF)**.** *For SGF on a loss function g, the value of the loss function g evolves according to the following SDE:*

$$dg(\beta(t)) = \left(-|\nabla g(\beta(t))|^2 + \frac{\eta^2}{2}\mathrm{tr}\left[\nabla^2 g(\beta(t)) \cdot \sigma(\beta(t))\sigma(\beta(t))^T\right]\right)dt - \eta\nabla g(\beta(t))^T\sigma(\beta(t))dW(t),$$

where $\Sigma(\beta) := \sigma(\beta)\sigma(\beta)^T$ is the covariance of the stochastic gradients at parameter value β. Also $\eta = \epsilon/\sqrt{m}$, where SGF approximates discrete time SGD with learning rate ϵ and batch size m. This can be written in a distributionally equivalent way as

$$dg(\beta(t)) = \left(-|\nabla g(\beta(t))|^2 + \frac{\eta^2}{2}\mathrm{tr}\left[\Sigma(\beta(t)) \cdot \nabla^2 g(\beta(t))\right]\right) dt - \eta\sqrt{\nabla g(\beta(t))^T \Sigma(\beta(t)) \nabla g(\beta(t))} dZ(t),$$

where $Z = \{Z(t)\}_{t\geq 0}$ is a 1-dimensional Brownian motion.

For the special case of least squares, let $r(t) = Y - X\beta(t)$ be the residual and define $M(t) = [XX^T \mathrm{diag}[r(t)^2]XX^T]/n^2$. Then we find that the equation for SGF with second moment matrix of stochastic gradients is

$$dg(\beta(t)) = \left(-\frac{\|X^T r(t)\|^2}{n^2} + \frac{\epsilon^2 \cdot \mathrm{tr}\left[M(t)\right]}{2m^2}\right) dt + \frac{\epsilon \cdot \|M(t)^{1/2} r(t)\|}{mn^{1/2}} dZ(t). \tag{S.4}$$

where $Z = \{Z(t)\}_{t\geq 0}$ is a 1-dimensional Brownian motion.

Proof. We start with the SDE for SGF

$$d\beta(t) = -\nabla g(\beta(t))dt + \eta\sigma(\beta(t))dW(t)$$

where $\sigma(\beta(t))\sigma(\beta(t))^T$ is the covariance of the gradients at parameter value $\beta(t)$. Then, Ito's rule leads to

$$\begin{aligned} dg(\beta(t)) &= \nabla g(\beta(t))^T d\beta(t) + \frac{1}{2} d\beta(t)^T \cdot H_x g \cdot d\beta(t) \\ &= \left(-|\nabla g(\beta(t))|^2 + \frac{\eta^2}{2}\mathrm{tr}\left[\sigma(\beta(t))^T \cdot \nabla^2 g(\beta(t)) \cdot \sigma(\beta(t))\right]\right) dt - \eta \nabla g(\beta(t))^T \sigma(\beta(t)) dW(t). \end{aligned}$$

For the special case of least squares loss, we have the following. We have already calculated most terms, and we have in addition that the second moment matrix of the gradients is

$$\sigma(\beta(t))\sigma(\beta(t))^T = \frac{1}{mn} X^T \mathrm{diag}[r(t)^2] X$$

where $r(t) = Y - X\beta(t)$ is the residual. Plugging in the terms for least squares,

$$dg(\beta(t)) = \left(-\frac{\|X^T r(t)\|^2}{n^2} + \frac{\epsilon^2 \cdot \mathrm{tr}\left[XX^T \mathrm{diag}[r(t)^2] XX^T\right]}{mn^2}\right) dt + \frac{\epsilon \cdot \|\mathrm{diag}[r(t)] XX^T r(t)\|}{m^{1/2} n^{3/2}} dZ(t).$$

Here $Z = \{Z(t)\}_{t\geq 0}$ is a 1-dimensional Brownian motion, which is obtained by transforming the original diffusion term, which is a linear combination of the entries of $dW(t)$, into a distributionally equivalent 1-dimensional process. Letting

$$M(t) = \frac{XX^T \mathrm{diag}[r(t)^2] XX^T}{n^2}$$

we can simplify the above as

$$dg(\beta(t)) = \left(-\frac{\|X^T r(t)\|^2}{n^2} + \frac{\epsilon^2 \cdot \mathrm{tr}\left[M(t)\right]}{m}\right) dt + \frac{\epsilon \cdot \|M(t)^{1/2} r(t)\|}{m^{1/2} n^{1/2}} dZ(t).$$

□

Comparing this with the noiseless case, i.e., when $\eta = 0$, we note that *both* the drift and the diffusion terms have changed. The drift term is reduced by a term that is proportional to η^2. The diffusion term is new altogether. This shows that for sufficiently large η, the drift will not be positive, and hence the process will not converge to a point mass limit distribution.

Let us start with studying the diffusion with the second moment matrix first. We will show a geometric contraction of the loss. We can bound the terms in the drift term as follows. We have (using $\odot$ for elementwise product of two conformable vectors or matrices)

$$r^2 \odot vec[\mathrm{diag}[(XX^T)^2]] \leq \|r\|^2 \max_i [(X^TX)^2]_{ii}$$
$$\|X^T r(t)\|^2 \geq \sigma_{\min}(X^T)^2 \|r(t)\|^2$$

The second inequality holds with $\sigma_{\min}(X^T)$ being the smallest nonzero singular value of X^T. Why? Because it is easy to see that we always have the decomposition

$$\frac{1}{2n}\|y - X\beta\|_2^2 = \frac{1}{2n}\|P_{\mathrm{col}(X)}y - X\beta\|_2^2 + \frac{1}{2n}\|P_{\mathrm{null}(X^T)}y\|_2^2, \tag{S.5}$$

i.e., we may think of the residual $r(t)$ above (and in the rest of this document) as $P_{\mathrm{col}(X)}(y - X\beta)$. It follows that $\|X^T r(t)\|_2^2 \geq s \cdot \|r(t)\|_2^2$, where s denotes the smallest nonzero singular value of X^T. It is also clear that the expressions for the gradient flow and stochastic gradient flow solutions do not change, if we use the decomposition in (S.5) as the loss. Hence, for the rest of this document, we always write $\sigma_{\min}$ to mean the smallest nonzero singular value. Below, we consider the situation when $p > n$ separately from the case $n \geq p$.

Now, let $V_t = g(\beta(t))$. Since V_t is an integral with respect to Brownian motion, it is a local martingale. Moreover, since it is non-negative (as the loss is non-negative), it follows that it is a supermartingale. Hence, writing (S.4) abstractly (defining A_t, B_t appropriately and denoting $Z(t) = Z_t$) as

$$dV_t = A_t dt + B_t dZ_t$$
$$V_T = \int_0^T A_t dt + \int_0^T B_t dZ_t$$

We have that $|B_t| = \epsilon \cdot \|\mathrm{diag}[r(t)]XX^T r(t)\|/(mn^{3/2}) \leq C\|r(t)\|^2 \leq C' V_t$ for some problem-dependent constants C, C'. Hence, as the growth of the diffusion coefficient is at most linear in $|V_t|$, it follows that V_t is not only a supermartingale, but also a martingale. Thus the expectation of the integral with respect to Brownian motion vanishes.

Below we will prove inequalities of the form $A_t \leq -A(V_t - B)$ for certain constants $A > 0$ and B. This leads to

$$\mathbb{E}V_T = \int_0^T \mathbb{E}A_t dt \leq A \int_0^T (\mathbb{E}V_t - B)dt$$

Letting $l(t) := \mathbb{E}V_t$ and differentiating the above leads to $l'(t) \leq -A(l(t) - B)$. We will see this in the overparametrized and underparametrized regimes in turn.

S.0.1. Overparametrized case

We find, with $M(t) = [XX^T \mathrm{diag}[r(t)^2]XX^T]/n^2$,

$$\left(-\frac{\|X^T r(t)\|^2}{n^2} + \frac{\epsilon^2 \cdot \mathrm{tr}\,[M(t)]}{2m^2}\right) \leq -\alpha[\|r(t)\|^2/2n] = -\alpha \cdot l(t)$$
$$\alpha = 2n[\sigma_{\min}(X)^2/n^2 - \epsilon^2 \max_i [(X^TX)^2]_{ii}/(2m^2)].$$

Then by taking expectations in the SDE for the loss, we find $l'(t) \leq -\alpha l(t)$. Hence $l(t) \leq \exp(-\alpha t) l_0$. For the diffusion with the covariance matrix of the gradients as a diffusion term, $\Sigma(\beta(t)) \prec \mathbb{E}\beta(t)\beta(t)^T$, hence $\mathrm{tr}\,[\Sigma(\beta(t)) \cdot \nabla^2 g(\beta(t))] \leq \mathrm{tr}\,[\mathbb{E}\beta(t)\beta(t)^T \cdot \nabla^2 g(\beta(t))]$. Thus, the drift term in this case is at most as large as the one in the second moment case, and so the contraction happens at least as fast. This proves the claim for the covariance matrix of the gradients as a diffusion term. The same argument will also apply to this case when $p < n$.

S.0.2. Underparametrized case

Now, if $p < n$, then in this case, in general the loss cannot converge to zero, because the number of equations is larger than the number of constraints. Instead, the loss converges close to the loss of the OLS estimator:

$$l^* = \|Y - X\hat{\beta}^{ols}\|^2/(2n) = \|P_X^{\perp} Y\|^2/(2n).$$

Then we can write

$$l(t) - l^* = \|X(\beta(t) - \hat{\beta}^{ols})\|^2/(2n)$$

Moreover, $X^T r(t) = -X^T X(\beta(t) - \hat{\beta}^{ols})$. Also, letting $b = P_X^{\perp} Y$, $r(t) = b + X(\hat{\beta}^{ols} - \beta(t))$, and hence

$$\|r(t)\|^2 \leq 2(\|b\|^2 + \|X(\hat{\beta}^{ols} - \beta(t))\|^2)$$

so that

$$r^2 \odot \operatorname{diag}[(XX^T)^2] \leq \|r\|^2 \max_i [(X^T X)^2]_{ii} \leq v + 2\|X(\hat{\beta}^{ols} - \beta(t))\|^2 \max_i [(X^T X)^2]_{ii}$$

where $v = 2\|b\|^2 \max_i [(X^T X)^2]_{ii}$. Let $q(t) = X(\beta(t) - \hat{\beta}^{ols})$. Then, as $l(t) - l^* = \|q(t)\|^2/(2n)$

$$\left(-\frac{\|X^T r(t)\|^2}{n^2} + \frac{\epsilon^2 \cdot \operatorname{tr}[M(t)]}{m} \right) \leq -c_0[l(t) - l^*] + C_0$$
$$c_0 = 2n[\sigma_{\min}(X)^2/n^2 - 2\epsilon^2 \max_i [(X^T X)^2]_{ii}/(2m^2)]$$
$$C_0 = 2\epsilon^2 \|b\|^2 \max_i [(X^T X)^2]_{ii}/(2m^2)$$

By taking expectations in the SDE for the loss, we find

$$l'(t) \leq -c_0[l(t) - l^*] + C_0.$$

Or also

$$l'(t) \leq -c_0[l(t) - (l^* + C_0/c_0)].$$

This shows that l converges geometrically to the level $l^* + C_0/c_0$, which is higher than the minimum OLS loss. In this case, the additional fluctuations occur because of the inherent noise in the algorithm.

Calculations for the In-Sample Risks, for Theorem 2

For in-sample risk, we have the bias-variance decomposition

$$\begin{aligned}
\operatorname{Risk}(\hat{\beta}^{\mathrm{sgf}}(t); \beta_0) &= \|\mathbb{E}_{\eta,Z}(\hat{\beta}^{\mathrm{sgf}}(t)) - \beta_0\|^2_{\hat{\Sigma}} + \operatorname{tr}\left[\operatorname{Cov}_{\eta,Z}(\hat{\beta}^{\mathrm{sgf}}(t))\hat{\Sigma}\right] \\
&= \|\mathbb{E}_{\eta,Z}(\hat{\beta}^{\mathrm{sgf}}(t)) - \beta_0\|^2_{\hat{\Sigma}} + \operatorname{tr}\left[\operatorname{Cov}_{\eta}(\hat{\beta}^{\mathrm{gf}}(t))\hat{\Sigma}\right] + \mathbb{E}_{\eta}\operatorname{tr}\left[\operatorname{Cov}_{Z}(\hat{\beta}^{\mathrm{sgf}}(t) \,|\, \eta)\hat{\Sigma}\right],
\end{aligned}$$

where we write $\|x\|^2_A = x^T A x$.

Following the same logic as in the proof of Theorem 2, we see that

$$\begin{aligned}
\|\mathbb{E}_{\eta,Z}(\hat{\beta}^{\mathrm{sgf}}(t)) - \beta_0\|^2_{\hat{\Sigma}} &= \left(\operatorname{Bias}^{\mathrm{in}}(\hat{\beta}^{\mathrm{sgf}}(t); \beta_0)\right)^2 = \left(\operatorname{Bias}^{\mathrm{in}}(\hat{\beta}^{\mathrm{gf}}(t); \beta_0)\right)^2 \\
\operatorname{tr}\left[\operatorname{Cov}_{\eta}(\hat{\beta}^{\mathrm{gf}}(t))\hat{\Sigma}\right] &= \operatorname{Var}^{\mathrm{in}}(\hat{\beta}^{\mathrm{gf}}(t)) \\
\mathbb{E}_{\eta}\operatorname{tr}\left[\operatorname{Cov}_{Z}(\hat{\beta}^{\mathrm{sgf}}(t) \,|\, \eta)\hat{\Sigma}\right] &\leq \epsilon \cdot \frac{n}{m} \sum_{i=1}^{p} \left(\frac{\tilde{w} s_i}{s_i - u/2} \big(\exp(-ut) - \exp(-2ts_i) \big) + \tilde{v}\big(1 - \exp(-2ts_i)\big) \right) s_i,
\end{aligned}$$

where (cf. Lemma 5 in Ali et al. (2018))

$$\left(\operatorname{Bias}^{\mathrm{in}}(\hat{\beta}^{\mathrm{gf}}(t); \beta_0)\right)^2 = \sum_{i=1}^{p} (v_i^T \beta_0)^2 s_i \exp(-2ts_i)$$
$$\operatorname{Var}^{\mathrm{in}}(\hat{\beta}^{\mathrm{gf}}(t)) = \frac{\sigma^2}{n} \sum_{i=1}^{p} (1 - \exp(-ts_i)).^2$$

Proof of Lemma 6

Looking back at (17), we have

$$\begin{aligned}
&\mathrm{Risk}(\hat{\beta}^{\mathrm{sgf}}(t);\beta_0) \qquad\qquad \text{(S.6)}\\
&\le \mathrm{Bias}^2(\hat{\beta}^{\mathrm{gf}}(t);\beta_0) + \mathrm{Var}_\eta(\hat{\beta}^{\mathrm{gf}}(t)) + \epsilon\cdot\frac{n}{m}\sum_{i=1}^p\Big(\frac{\tilde{w}s_i}{s_i-u/2}\big(\exp(-ut)-\exp(-2ts_i)\big)+\tilde{v}\big(1-\exp(-2ts_i)\big)\Big)\\
&= \mathrm{Bias}^2(\hat{\beta}^{\mathrm{gf}}(t);\beta_0) + \mathrm{Var}_\eta(\hat{\beta}^{\mathrm{gf}}(t)) + T + \epsilon\cdot\frac{n}{m}\sum_{i=1}^p\tilde{v}\big(1-\exp(-2ts_i)\big),
\end{aligned}$$

where we let

$$T = \epsilon\cdot\frac{n}{m}\sum_{i=1}^p\Big(\frac{\tilde{w}s_i}{s_i-u/2}\big(\exp(-ut)-\exp(-2ts_i)\big)\Big).$$

Focusing on just T for now, and noting that $s_i > u/2$ for $i = 1,\ldots,p$, we see

$$T \le p\tilde{w}\epsilon\cdot\frac{n}{m}\frac{\mu}{\mu-u/2}\exp(-ut),$$

implying that

$$(T/\alpha)^{2L/u} \le \exp(-2Lt) \le \exp(-2s_it),$$

for $i = 1,\ldots,p$, which means that

$$\|\beta_0\|_2^2\cdot(T/\alpha)^{2L/u} = \|V\beta_0\|_2^2\cdot(T/\alpha)^{2L/u} \le \sum_{i=1}^p(v_i^T\beta_0)^2\exp(-2s_it) = \mathrm{Bias}^2(\hat{\beta}^{\mathrm{gf}}(t);\beta_0).$$

Here, we used the eigendecomposition $\hat{\Sigma} = VSV^T$, and Lemma 5 in Ali et al. (2018). Hence,

$$\begin{aligned}
T &= \epsilon\cdot\frac{n}{m}\sum_{i=1}^p\Big(\frac{\tilde{w}s_i}{s_i-u/2}\big(\exp(-ut)-\exp(-2ts_i)\big)\Big)\\
&\le \alpha\cdot\left(\frac{|\mathrm{Bias}(\hat{\beta}^{\mathrm{gf}}(t);\beta_0)|}{\|\beta_0\|_2}\right)^{\mu/L}\\
&= \delta\cdot|\mathrm{Bias}(\hat{\beta}^{\mathrm{gf}}(t);\beta_0)|^{1/\kappa}. \qquad\qquad \text{(S.7)}
\end{aligned}$$

Therefore, putting (S.6) and (S.7) together, along with Lemma 5 in Ali et al. (2018), we obtain

$$\begin{aligned}
&\mathrm{Risk}(\hat{\beta}^{\mathrm{sgf}}(t);\beta_0)\\
&\le \mathrm{Bias}^2(\hat{\beta}^{\mathrm{gf}}(t);\beta_0) + \delta\cdot|\mathrm{Bias}(\hat{\beta}^{\mathrm{gf}}(t);\beta_0)|^{1/\kappa} + \underbrace{\frac{\sigma^2}{n}\sum_{i:s_i>0}\frac{(1-\exp(-ts_i))^2}{s_i}}_{A} + \underbrace{\tilde{v}\epsilon\cdot\frac{n}{m}\sum_{i:s_i>0}(1-\exp(-2ts_i))}_{B}.
\end{aligned}$$

Now for convenience, write $A = \frac{\sigma^2}{n}\sum_{i:s_i>0}a_i$ and $B = \frac{\epsilon}{2m}\cdot\sum_{i:s_i>0}b_i$. Let $f(x)$ be the continuous extension of $\tilde{f}(x)$, where $\tilde{f}(x) = x\frac{1+\exp(-x)}{1-\exp(-x)}$ (i.e., $f(x) = \tilde{f}(x)$ when $x > 0$, but $f(x) = 2$ when $x = 0$). It can be checked that $f(x)$ is nondecreasing, so that $\sup_{x\in[0,L]}f(tx) = f(tL)$.

As $f(ts_i) \le f(tL)$, we have for each i such that $s_i > 0$,

$$ts_i\frac{1+\exp(-ts_i)}{1-\exp(-ts_i)} \le f(tL).$$

Multiplying both sides by $(1-\exp(-ts_i))^2$ and rearranging yields

$$(1+\exp(-ts_i))(1-\exp(-ts_i)) \le \frac{f(tL)}{t}\frac{(1-\exp(-ts_i))^2}{s_i},$$

i.e.,

$$b_i \leq \frac{f(tL)}{t} a_i.$$

Therefore,

$$A + B \leq \left(1 + \frac{\tilde{v}\epsilon \cdot n^2 f(tL)}{mt\sigma^2}\right) A.$$

Now, note that $f(x)$ is increasing on $x > 0$, and $f(x)/x$ is decreasing on $x > 0$. Also, note that $f(x) \leq 2.164$ when $x \leq 1$, and $f(x)/x \leq 2.164$ when $x > 1$. Thus, $f(x)/x \leq \max\{2.164/x, 2.164\}$. So,

$$A + B \leq \left(1 + 2.164\epsilon \cdot \frac{\tilde{v}n^2 \max(1/t, L)}{m\sigma^2}\right) A.$$

Putting together the pieces, we obtain

$$\mathrm{Risk}(\hat{\beta}^{\mathrm{sgf}}(t); \beta_0) \leq \mathrm{Bias}^2(\hat{\beta}^{\mathrm{gf}}(t); \beta_0) + \delta \cdot |\mathrm{Bias}(\hat{\beta}^{\mathrm{gf}}(t); \beta_0)|^{1/\kappa} + \gamma(t) \cdot \mathrm{Var}(\hat{\beta}^{\mathrm{gf}}(t)),$$

which shows the claim for gradient flow. Applying Theorem 1 shows the result for ridge. Turning to in-sample risk, the exact same bounds actually follow by similar arguments, just as discussed before. □

Proof of Lemma 7

Letting $X = n^{1/2} U S^{1/2} V^T$ be a singular value decomposition, we may express

$$\hat{\beta}^{\mathrm{gf}}(t) = n^{-1/2} V S^+ (I - \exp(-tS)) S^{1/2} U^T y = n^{-1/2} t^{1/2} V (tS)^+ (I - \exp(-tS)) (tS)^{1/2} U^T y,$$

and

$$\hat{\beta}^{\mathrm{ridge}}(1/t) = n^{-1/2} V \Big(S + \frac{1}{t} \cdot I\Big)^{-1} S^{1/2} U^T y = n^{-1/2} t^{1/2} V (tS + I)^{-1} (tS)^{1/2} U^T y.$$

Therefore,

$$\|\hat{\beta}^{\mathrm{gf}}(t) - \hat{\beta}^{\mathrm{ridge}}(1/t)\|_2 = \Big\| n^{-1/2} t^{1/2} \Big((tS)^+ (I - \exp(-tS)) - (tS + I)^{-1} \Big) (tS)^{1/2} U^T y \Big\|_2. \tag{S.8}$$

Now define $f(x) = (1 - \exp(-x))(1 + x)/x$ with domain $x \geq 0$ (let $f(0) = 0$). Lemma 7 in Ali et al. (2018) shows that f attains its unique maximum at $x^* = 1.7933$, where $f(x^*) = 1.2985$. Moreover, it can be checked that f is unimodal. This means that, for $i = 1, \ldots, p$ and $t \leq 1.7933/L$,

$$\frac{(1 - \exp(-s_i t))(1 + s_i t)}{s_i t} \leq \frac{(1 - \exp(-Lt))(1 + Lt)}{Lt},$$

i.e.,

$$\frac{1 - \exp(-s_i t)}{s_i t} - \frac{1}{1 + s_i t} \leq (g(t) - 1) \cdot \frac{1}{1 + s_i t}.$$

Similar reasoning shows that, for $t \geq 1.7933/\mu$,

$$\frac{1 - \exp(-s_i t)}{s_i t} - \frac{1}{1 + s_i t} \leq (g(t) - 1) \cdot \frac{1}{1 + s_i t}.$$

When $1.7933/L < t < 1.7933/\mu$, we may simply take

$$\frac{1 - \exp(-s_i t)}{s_i t} - \frac{1}{1 + s_i t} \leq (1.2985 - 1) \cdot \frac{1}{1 + s_i t} = (g(t) - 1) \cdot \frac{1}{1 + s_i t}.$$

Returning to (S.8), we have shown that

$$(tS)^+ (I - \exp(-tS)) - (tS + I)^{-1} \preceq (g(t) - 1) \cdot (tS + I)^{-1}.$$

Thus,

$$\begin{aligned}\|\hat{\beta}^{\mathrm{gf}}(t) - \hat{\beta}^{\mathrm{ridge}}(1/t)\|_2 &\leq \|(g(t)-1)\cdot n^{-1/2}t^{1/2}(tS+I)^{-1}(tS)^{1/2}U^T y\|_2 \\ &= (g(t)-1)\cdot\|\hat{\beta}^{\mathrm{ridge}}(1/t)\|_2,\end{aligned}$$

as claimed. □

Additional Numerical Simulations

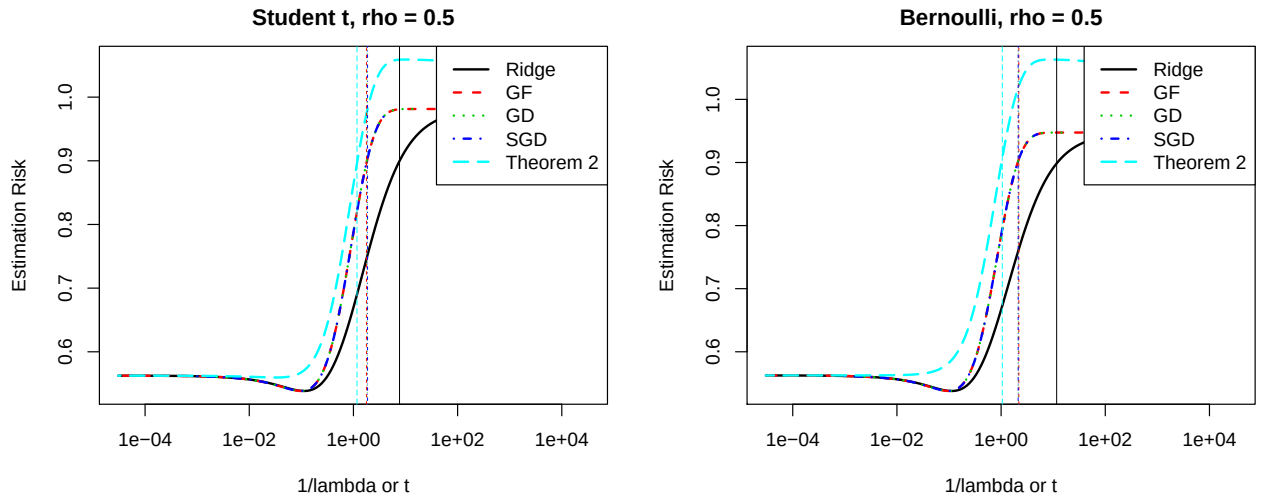

Figure S.1. Risks for ridge regression, discrete-time SGD, stochastic gradient flow (as in Theorem 2), discrete-time gradient descent, and gradient flow on Student-t and Bernoulli data, where $n = 100$, $p = 500$, $m = 20$, and ϵ was set following Lemma 5. The excess risk of stochastic gradient flow over ridge is given by the distance between the cyan and black curves. The vertical lines show the stopping times that balance bias and variance.

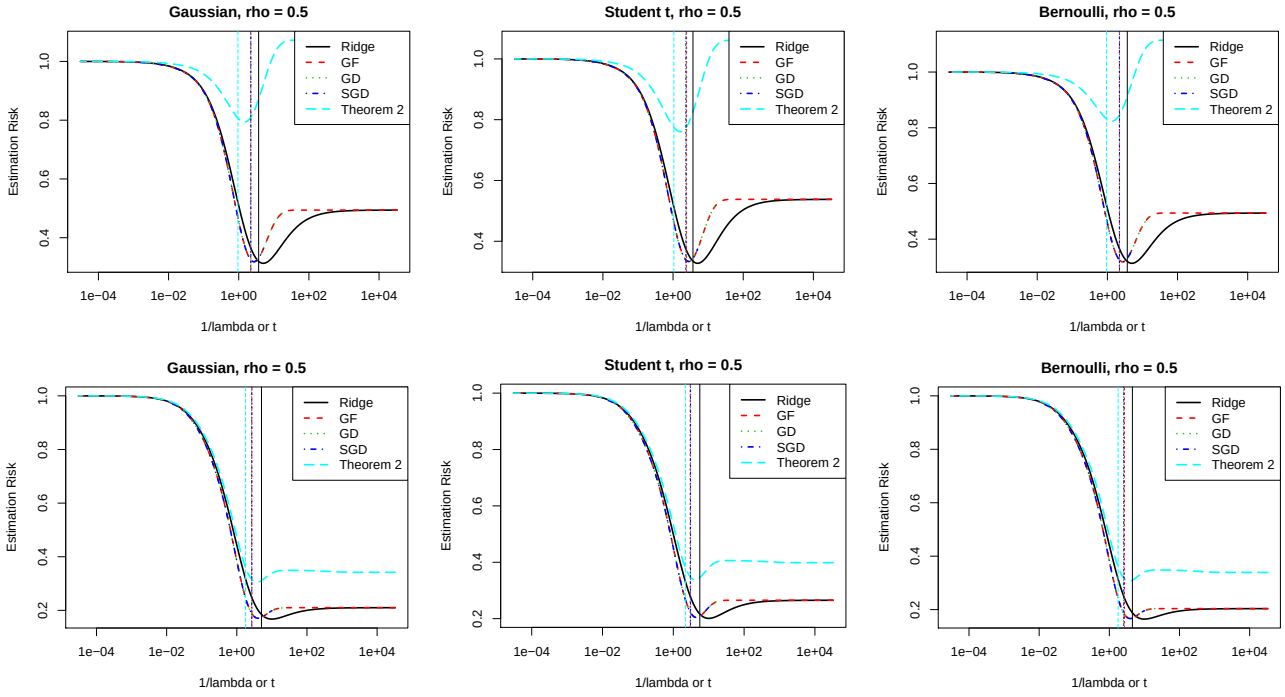

Figure S.2. Risks for ridge regression, discrete-time SGD, stochastic gradient flow (as in Theorem 2), discrete-time gradient descent, and gradient flow on Gaussian, Student-t, and Bernoulli data. The excess risk of stochastic gradient flow over ridge is given by the distance between the cyan and black curves. The vertical lines show the stopping times that balance bias and variance. In the first row, we set $n = 500$, $p = 100$, $m = 20$, and ϵ following Lemma 5. In the second, we set $n = 100$, $p = 10$, $m = 10$, and ϵ following Lemma 5.

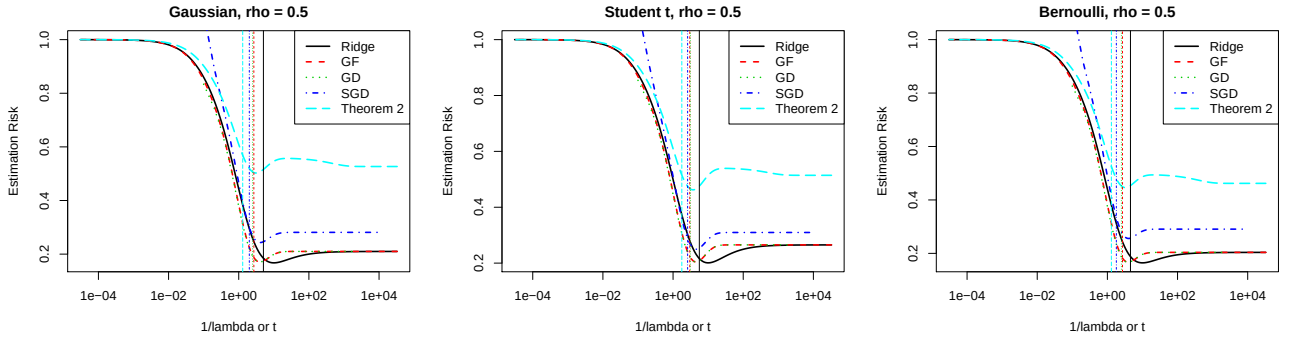

Figure S.3. Risks for ridge regression, discrete-time SGD, stochastic gradient flow (as in Theorem 2), discrete-time gradient descent, and gradient flow on Gaussian, Student-t, and Bernoulli data, where $n = 100$, $p = 10$, $m = 2$, and $\epsilon = 0.1$. The excess risk of stochastic gradient flow over ridge is given by the distance between the cyan and black curves. The vertical lines show the stopping times that balance bias and variance.

Further Comparisons Between the Non-Constant and Constant Covariance Processes

Following the discussion surrounding Lemma 3 in the main paper, it is interesting to compare the excess risk bound (18) to the analogous bound for the time-homogeneous process (10). Using Lemma 2, we may denote the solution to the time-homogeneous process (10) as $\tilde{\beta}^{\mathrm{sgf}}(t)$. Then, following the same logic as in the proof of Theorem 2, we obtain

$$\mathrm{Risk}(\tilde{\beta}^{\mathrm{sgf}}(t); \beta_0) = \mathrm{Bias}^2(\hat{\beta}^{\mathrm{gf}}(t); \beta_0) + \mathrm{Var}_\eta(\hat{\beta}^{\mathrm{gf}}(t)) + \mathrm{tr}\,\mathbb{E}_\eta\big[\mathrm{Cov}_Z(\tilde{\beta}^{\mathrm{sgf}}(t) \,|\, \eta)\big].$$

Ito's isometry shows that

$$\begin{aligned}
\mathrm{tr}\,\mathbb{E}_\eta\big[\mathrm{Cov}_Z(\tilde{\beta}^{\mathrm{sgf}}(t) \,|\, \eta)\big] &= \frac{\epsilon}{m} \cdot \mathrm{tr}\,\mathbb{E}_\eta\left[\left(\exp(-t\hat{\Sigma})\int_0^t \exp(\tau\hat{\Sigma})\hat{\Sigma}^{1/2} dW(\tau)\right)\left(\exp(-t\hat{\Sigma})\int_0^t \exp(\tau\hat{\Sigma})\hat{\Sigma}^{1/2} dW(\tau)\right)^T\right] \\
&= \frac{\epsilon}{m} \cdot \mathrm{tr}\,\mathbb{E}_\eta\left[\left(\int_0^t \exp(-t\hat{\Sigma})\exp(\tau\hat{\Sigma})\hat{\Sigma}\exp(\tau\hat{\Sigma})\exp(-t\hat{\Sigma}) d\tau\right]\right. \\
&= \frac{\epsilon}{m} \cdot \mathrm{tr}\left(\hat{\Sigma}\int_0^t \exp(2(\tau - t)\hat{\Sigma}) d\tau\right) \qquad \text{(S.9)} \\
&= \frac{\epsilon}{2m} \cdot \mathrm{tr}\left(\hat{\Sigma}\hat{\Sigma}^+(I - \exp(-2t\hat{\Sigma}))\right).
\end{aligned}$$

Finally, expanding $\exp(-2t\hat{\Sigma})$ into its power series representation and using the eigendecomposition $\hat{\Sigma} = VSV^T$ shows that $\hat{\Sigma}\hat{\Sigma}^+(I - \exp(-2t\hat{\Sigma})) = I - \exp(-2t\hat{\Sigma})$, which gives

$$\mathrm{Risk}(\tilde{\beta}^{\mathrm{sgf}}(t); \beta_0) = \mathrm{Risk}(\hat{\beta}^{\mathrm{gf}}(t); \beta_0) + \frac{\epsilon}{2m} \cdot \mathrm{tr}\left(I - \exp(-2t\hat{\Sigma})\right). \qquad \text{(S.10)}$$

From (S.10), it is straightforward to derive expressions analogous to those appearing in (12), (13), (18), for the process $\tilde{\beta}^{\mathrm{sgf}}(t)$. It is also possible to follow the same logic as in the proof of Lemma 6 to arrive at a similar expression for $\tilde{\beta}^{\mathrm{sgf}}(t)$ (i.e., with $\gamma(t)$ suitably redefined).

Comparing the preceding calculations with those given in the proof of Theorem 2, we see that a key simplification occurs in (S.9), above. Here, the (relatively) complicated expression appearing in the proof of Theorem 2,

$$\frac{2n\epsilon}{m} \cdot \int_0^t \mathbb{E}_\eta[f(\hat{\beta}^{\mathrm{sgf}}(\tau))] d\tau,$$

is replaced with the comparatively simpler expression $(\epsilon/m) \cdot \hat{\Sigma}$ in (S.9). This simplification allows the risk expression in (S.10) to hold with equality, though it is evidently less refined than the bound appearing in, e.g., (12).

References

Ali, A., Kolter, J. Z., and Tibshirani, R. J. A continuous-time view of early stopping for least squares regression. *arXiv preprint arXiv:1810.10082*, 2018.

Hoffer, E., Hubara, I., and Soudry, D. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*, pp. 1731–1741, 2017.

Hu, W., Li, C. J., Li, L., and Liu, J.-G. On the diffusion approximation of nonconvex stochastic gradient descent. *arXiv preprint arXiv:1705.07562*, 2017.

Nguyen, V. A., Shafieezadeh-Abadeh, S., Kuhn, D., and Esfahani, P. M. Bridging bayesian and minimax mean square error estimation via wasserstein distributionally robust optimization. *arXiv preprint arXiv:1911.03539*, 2019.

Øksendal, B. *Stochastic differential equations*. Springer, 2003.

Zhang, C., Kjellstrom, H., and Mandt, S. Determinantal point processes for mini-batch diversification. *arXiv preprint arXiv:1705.00607*, 2017.

Zhu, Z., Wu, J., Yu, B., Wu, L., and Ma, J. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from minima and regularization effects. *arXiv preprint arXiv:1803.00195*, 2018.