Appendix

Organisation of the Appendix. The Appendix is structured as follows. In Section A, we give more precisions regarding the way we model stochastic gradient descent as a stochastic gradient flow. Section B.6 is the core of the Appendix as it provides the proof of the theorem in a self-contained fashion. For the sake of completeness, in Section C we gather the results on the link between mirror-descent and implicit bias as well as give convergence results in the deterministic case (gradient flow). In Section D.1, we provide more experiments supporting our results. In Section E.2, we discuss some extensions of our results ; (E.1) regarding a more general stochastic gradient flow model and in (E.2) we extend our results to depths $p \geq 3$. Finally, Section F provides the technical material needed for the proofs of our results.

A Details on the SDE modelling

We recall that the SGD recursion writes for $t \geqslant 1$ as:

$$\begin{aligned} w_{t+1,+} &= w_{t,+} - \gamma \langle \beta_w - \beta^*, x_{i_t} \rangle \, x_{i_t} \odot w_{t,+} \\ w_{t+1,-} &= w_{t,-} + \gamma \langle \beta_w - \beta^*, x_{i_t} \rangle \, x_{i_t} \odot w_{t,-} \end{aligned} \qquad \text{where } i_t \sim \mathrm{Unif}(1,n).$$

Since the full gradient is $\nabla_{w_\pm} L(w) = \pm \big[\frac{1}{n} \sum_{k=1}^n \langle \beta_w - \beta^*, x_k \rangle \, x_k \big] \odot w_\pm \in \mathbb{R}^d$. We can rewrite the recursion as:

$$w_{t+1,\pm} = w_{t,\pm} - \gamma \nabla_{w_\pm} L(w_t) \mp \gamma \Big[\langle \beta_{w_t} - \beta^*, x_{i_t} \rangle \, x_{i_t} - \frac{1}{n} \sum_{k=1}^n \langle \beta_{w_t} - \beta^*, x_k \rangle \, x_k \Big] \odot w_{t,\pm}.$$

Now notice that

$$\langle \beta - \beta^*, x_{i_t} \rangle \, x_{i_t} - \frac{1}{n} \sum_{k=1}^n \langle \beta - \beta^*, x_k \rangle \, x_k = X^\top \Big(\langle \beta - \beta^*, x_{i_t} \rangle \mathbf{e}_{i_t} - \mathbb{E}_{i_t} \big[\langle \beta - \beta^*, x_{i_t} \rangle \mathbf{e}_{i_t} \big] \Big),$$

where $\mathbf{e}_i$ is the i^{th} element of the $\mathbb{R}^n$-canonical basis. Let us denote by $\xi_{i_t}(\beta) = -\big(\langle \beta - \beta^*, x_{i_t} \rangle \mathbf{e}_{i_t} - \mathbb{E}_{i_t} \big[\langle \beta - \beta^*, x_{i_t} \rangle \mathbf{e}_{i_t} \big] \big)$. It is a zero-mean random variable with values in $\mathbb{R}^n$ and it can be seen as a multiplicative noise, i.e., proportional to $\beta - \beta^*$, which vanishes at the optimum. The SGD recursion then writes as:

$$\begin{aligned} w_{t+1,\pm} &= w_{t,\pm} - \gamma \nabla_{w_\pm} L(w_t) \pm \gamma \big[X^\top \xi_{i_t}(\beta_t) \big] \odot w_{t,\pm} \\ &= w_{t,\pm} - \gamma \nabla_{w_\pm} L(w_t) \pm \gamma \, \mathrm{diag}(w_{t,\pm}) X^\top \xi_{i_t}(\beta_t). \end{aligned}$$

As we are interested in the stochastic differential model of the SGD recursion, let us now compute the covariance of the SGD noise. We first notice that

$$\begin{aligned} \mathrm{Cov}_{i_t}[\xi_{i_t}(\beta)] &= \mathbb{E}_{i_t}[\xi_{i_t}(\beta)^{\otimes 2}] \\ &= \mathbb{E}_{i_t}[(\langle \beta - \beta^*, x_{i_t} \rangle \mathbf{e}_{i_t})^{\otimes 2}] - \mathbb{E}_{i_t}[\langle \beta - \beta^*, x_{i_t} \rangle \mathbf{e}_{i_t}]^{\otimes 2} \\ &= \frac{1}{n} \begin{pmatrix} \langle \beta - \beta^*, x_1 \rangle^2 & & 0 \\ & \ddots & \\ 0 & & \langle \beta - \beta^*, x_n \rangle^2 \end{pmatrix} - \frac{1}{n^2} \Big(\langle \beta - \beta^*, x_i \rangle \langle \beta - \beta^*, x_j \rangle \Big)_{1 \leq i,j \leq n} \\ &= \frac{4}{n} \begin{pmatrix} L_1(\beta) & & 0 \\ & \ddots & \\ 0 & & L_n(\beta) \end{pmatrix} - \frac{1}{n^2} \Big(\langle \beta - \beta^*, x_i \rangle \langle \beta - \beta^*, x_j \rangle \Big)_{1 \leq i,j \leq n} \end{aligned}$$

where $L_i(\beta) = \frac{1}{4} \langle \beta - \beta^*, x_i \rangle^2$ is the individual loss of the observation x_i, such that $L(\beta) = \frac{1}{n} \sum_{i=1}^n L_i(\beta)$.

Thus, the covariance satisfies the relation $\mathrm{Cov}_{i_t}[\xi_{i_t}(\beta)] = \frac{4}{n} \mathrm{diag}(L_i(\beta))_{1 \leq i \leq n} + O(\frac{1}{n^2})$. From this expression we can obtain a good model for $\mathrm{Cov}_{i_t}[\xi_{i_t}(\beta)]$. First, we neglect the second term of order $1/n^2$. Then, we assume that all partial losses are approximately uniformly equal to their mean: i.e. for any i, $L_i(\beta) \cong \mathbb{E}_{i_t}[L_{i_t}(\beta)]$ (the general case is discussed Appendix E.1). Hence,

$$\mathrm{Cov}_{i_t}[\xi_{i_t}(\beta)] \cong \frac{4}{n} \, \mathrm{diag} \Big(\frac{1}{n} \sum_i L_i(\beta) \Big) = \frac{4}{n} L(\beta) I_n.$$