A Second-Order Method for Stochastic Bandit Convex Optimisation

Tor Lattimore LATTIMORE@GOOGLE.COM
Google DeepMind, London

András György AGYORGY@GOOGLE.COM
Google DeepMind, London

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We introduce a simple and efficient algorithm for unconstrained zeroth-order stochastic convex bandits and prove its regret is at most $(1 + r/d)[d^{1.5}\sqrt{n} + d^3] \operatorname{polylog}(n, d, r)$ where n is the horizon, d the dimension and r is the radius of a known ball containing the minimiser of the loss.
Keywords: Bandits; zeroth-order convex optimisation.

1. Introduction

Let $\|\cdot\|$ be the standard Euclidean norm and $f : \mathbb{R}^d \to \mathbb{R}$ be a convex loss function and assume that

(a) f is Lipschitz: $f(x) - f(y) \leq \|x - y\|$ for all $x, y \in \mathbb{R}^d$;

(b) there exists an $x_\star \in \mathbb{R}^d$ such that $f(x_\star) = \inf_{x \in \mathbb{R}^d} f(x)$;

(c) the learner has access to a constant $r \geq 1$ and initial point $x_\circ$ such that $\|x_\star - x_\circ\| \leq r$.

A learner interacts with an environment over n rounds. In each round t the learner chooses $X_t \in \mathbb{R}^d$ and observes $Y_t = f(X_t) + \varepsilon_t$ where $(\varepsilon_t)_{t=1}^n$ is a sequence of conditionally zero mean subgaussian random variables (precise condition given in Equation (1) below). As usual in bandit problems, X_t is only allowed to depend on previous observations $X_1, Y_1, \ldots, X_{t-1}, Y_{t-1}$ and possibly an external source of randomness. Our focus is on the cumulative regret $\operatorname{Reg}_n = \sum_{t=1}^n f(X_t) - f(x_\star)$, that is, how much more loss the algorithm suffers compared to the optimal action $x_\star$ in n rounds. The main contribution is the following regret guarantee for a simple algorithm for which the computation per round is dominated by finding the eigendecomposition of a $d \times d$ matrix.

Theorem 1 *With probability at least $1 - 6/n$, the regret of Algorithm 1 is upper bound by*

$$Reg_n \leq \operatorname{const} \left(1 + \frac{r}{d}\right) \left[d^{1.5}\sqrt{n} + d^3\right] (1 + \log\max(n, d, r))^4 ,$$

where const *is a universal constant.*

The best known bound in this setting is $\mathbb{E}[\operatorname{Reg}_n] \leq r d^{2.5}\sqrt{n} \operatorname{polylog}(n, d, r)$, which does not come with an efficient algorithm (Lattimore, 2020). Our high-level idea is to combine online Newton step (Hazan et al., 2007) with randomised estimators of the gradient and Hessian of the surrogate loss function used by Bubeck et al. (2017) and Lattimore and György (2021). Although our analysis and algorithm are designed for the regret setting, an important consequence is an improved bound for stochastic zeroth-order convex optimisation.

Corollary 2 *For $\varepsilon > 0$ and $n \geq \left\lceil \text{const}\left(\frac{d^3}{\varepsilon^2} + \frac{d^3}{\varepsilon}\right)\left(1 + \frac{r}{d}\right)^2 (1 + \log(\max(1/\varepsilon, d, r)))^8 \right\rceil$,*

$$\mathbb{P}\left(f\left(\frac{1}{n}\sum_{t=1}^n X_t\right) \geq f(x_\star) + \varepsilon\right) \leq \frac{6}{n}\,,$$

where $X_1, \ldots, X_n$ are the actions chosen by Algorithm 1.

Until now, the best known bound on the sample complexity of an efficient algorithm in this setting was $\tilde{O}(\frac{d^{7.5}}{\varepsilon^2})$ by Belloni et al. (2015). Lattimore (2020) demonstrated the existence of a procedure for which the sample complexity is at most $\tilde{O}(\frac{d^5}{\varepsilon^2})$, but the approach is non-constructive. We emphasise that both of these works are intended for the harder constrained setting.

Related work There is an ever-growing literature on convex bandits in a variety of settings. Our setup is unusual because there are no constraints on the domain of the function to be optimised. Of course, algorithms that handle constraints can be used in our setup because of the assumption that the minimum lies in a known ball. The other direction is not clear. We expect that suitable modifications of our ideas will lead to algorithms for the constrained case, but not without effort, ingenuity and possibly some dimension-dependent cost. More on this in the discussion.

The most natural idea to extend the standard machinery for stochastic gradient descent to the zeroth-order bandit setting is to use importance-weighted gradient estimators of a smoothed approximation of f, which was the approach taken by Kleinberg (2005), Flaxman et al. (2005) and Saha and Tewari (2011). This leads to simple generalisations of gradient descent that are straightforward to implement and analyse. Sadly this approach does not lead to $\sqrt{n}$ regret without strong convexity.

In the stochastic setting it is possible to adapt tools from the classical zeroth-order optimisation literature as was shown by Agarwal et al. (2013), who proved $\text{poly}(d)\sqrt{n}$ regret for constrained stochastic convex bandits without smoothness or strong convexity assumptions. These ideas were improved by Lattimore and György (2021) leading to a better dimension dependence. Bubeck et al. (2015) showed that $\sqrt{n}$ regret is also possible without strong convexity/smoothness in the adversarial setting when $d = 1$. Their approach non-constructively leveraged the information-theoretic machinery of Russo and Van Roy (2014) and did not yield an algorithm. There followed a flurry of results generalising this to higher dimensions and/or polynomial time algorithms (Hazan and Li, 2016; Bubeck et al., 2017; Bubeck and Eldan, 2018; Lattimore, 2020). None of these algorithms is particularly straightforward to implement.

What is missing in the literature is a simple algorithm with $\sqrt{n}$ regret in any setting without strong convexity. Interestingly, Hu et al. (2016) proved a negative result showing that any analysis that uses gradient estimators must use more properties of these estimators than any naive bias-variance decomposition that appeared in previous work (Kleinberg, 2005; Flaxman et al., 2005). This negative result does not hold in the strongly convex setting, where gradient-based methods give $\sqrt{n}$ regret (Agarwal et al., 2010; Hazan and Levy, 2014; Ito, 2020; Luo et al., 2022). Finally, Suggala et al. (2021) study the adversarial problem where the loss function is (nearly) quadratic. They design a computationally efficient algorithm with $d^{16}\sqrt{n}\,\text{polylog}(n)$ regret. Like us, they also use Hessian estimates to control a focus region. Because they study the adversarial setting the situation is more subtle. If the adversary dramatically changes the minimiser the algorithm needs to detect the change and restart or broaden the focus region. The current state of affairs is given in Table 1.

Authors	Constrained	Adversarial	Lipschitz	Smooth	Strongly convex	Regret *	Comp.
Flaxman et al. (2005)	✓	✓	✓			$\sqrt{d}n^{\frac{3}{4}}$	$O(d)$ †
Saha and Tewari (2011)	✓	✓		✓		$\nu^{\frac{1}{3}} d^{\frac{2}{3}} n^{\frac{2}{3}}$	$O(d)$ †
Hazan and Levy (2014)	✓	✓		✓	✓	$d\sqrt{\nu n}$	$O(d)$ †
Bubeck et al. (2017)	✓	✓				$d^{10.5}\sqrt{n}$	$\text{poly}(d, n)$
Lattimore (2020)	✓	✓				$d^{2.5}\sqrt{n}$	$\exp(d, n)$
Lattimore and György (2021)	✓					$d^{4.5}\sqrt{n}$	$O(d)$ ‡
This work			✓			$d^{1.5}\sqrt{n}$	$O(d^3)$

$\star$ All regret bounds hold up to logarithmic factors for sufficiently large n and omit any dependence on r or the range of losses. The parameter ν is the self-concordance parameter for a barrier on the constraint set, which is $\nu = 1$ for the ball and information-theoretically never more than d. † These computation bounds assume the constraint set is a ball. ‡ The algorithm uses the ellipsoid method and needs logarithmically many updates of $O(d^3)$ for the unconstrained case or $O(d^4)$ with a separation oracle on the constraint set.

Table 1: The current Pareto frontier for unconstrained zeroth-order bandit convex optimisation. Shaded cells correspond to poor behaviour of the corresponding algorithm in relation to the property associated with the cell. Algorithms that do not depend on a Lipschitz assumption assume the loss is bounded on the constraint set. The best lower bound is still $\Omega(d\sqrt{n})$ and uses linear functions (Dani et al., 2008). Algorithms for the constrained setting can also be used in the unconstrained one but not a-priori the other way.

Notation The vector of all zeros is $\mathbf{0}$ and the identity matrix is $\mathbb{1}$, which will always be d-dimensional. This should not be confused with the indicator function, denoted by $\mathbf{1}(\cdot)$. The density (with respect to Lebesgue) of the Gaussian distribution with mean μ and covariance Σ is $\mathcal{N}(\mu, \Sigma)$. Given vector x and square matrix A, $\|x\|$ is the standard Euclidean norm and $\|x\|_A^2 = x^\top Ax$. The operator norm of a real matrix A is $\|A\| = \max_{x\neq\mathbf{0}} \|Ax\|/\|x\|$. For positive semidefinite matrices A and B we write $A \leq B$ or $B \geq A$ to mean that $B - A$ is positive semidefinite. For random elements X and Y taking values in the same space we write $X \overset{d}{=} Y$ if $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ for all measurable A. The complement of an event E is E^c. For a real random variable W and $k \in \{1, 2\}$, let $\|W\|_{\psi_k} = \inf\{t > 0 : \mathbb{E}[\exp(|W|^k/t^k)] \leq 2\}$. A random variable is called subgaussian if $\|W\|_{\psi_2} < \infty$ and subexponential if $\|W\|_{\psi_1} < \infty$. A simple corollary of the definitions is that $\|W^2\|_{\psi_1} = \|W\|_{\psi_2}^2$.

Constants The parameters of our algorithm are defined in terms of absolute constants. We let C and c represent suitably large/small absolute positive constants and $\mathrm{P} = \max(2, n, d, r)^m$ where $m \geq 1$ is a suitably large absolute constant. One can always check in the proofs that an appropriate choice of these constants is possible, by first choosing a small enough c, then a large enough C, and finally a large enough exponent m.

Sigma-algebras and noise sequence Let $\mathscr{F}_t = \sigma(X_1, Y_1, \ldots, X_t, Y_t)$ be the σ-algebra generated by the action/loss sequence and let $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathscr{F}_t]$ and $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot|\mathscr{F}_t)$. Occasionally we need to specify the probability measure with respect to which an Orlicz norm is defined. Given a random variable X defined on some measurable space $(\Omega, \mathscr{F})$, we write $\|X\|_{\mathbb{P},\psi_k}$ for the corresponding Orlicz norm with respect to probability measure $\mathbb{P}$ on $(\Omega, \mathscr{F})$. Very often $\mathbb{P}$ is $\mathbb{P}_t$ and for this we make the abbreviation $\|X\|_{t,\psi_k} = \|X\|_{\mathbb{P}_t,\psi_k}$. Our assumption on the noise sequence $(\varepsilon_t)_{t=1}^n$ is that

$$\|\varepsilon_t\|_{\mathbb{P}_{t-1}(\cdot|X_t),\psi_2} \leq 1 \text{ and } \mathbb{E}_{t-1}[\varepsilon_t|X_t] = 0\,. \tag{1}$$

That is, after conditioning on the past action/losses and the current action, the noise is subgaussian and has mean zero. If the bound on the Orlicz norm in Equation (1) is replaced with the assumption that $\|\varepsilon_t\|_{\mathbb{P}_{t-1}(\cdot|X_t),\psi_2} \leq \sigma$, then Theorem 1 continues to hold with $1 + r/d$ replaced with $\sigma + r/d$.

2. Algorithm

Our algorithm, shown in Algorithm 1, is an instantiation of online Newton step (Hazan et al., 2007), which is a second-order method designed for the full information setting. In the zeroth-order setting considered here, the learner does not have access to the gradient or the Hessian. Nevertheless, our algorithm uses online Newton step, replacing the exact gradient and Hessian with stochastic estimates of a surrogate loss function rather than the actual loss function. Interestingly, online Newton step with quadratic losses is equivalent to continuous exponential weights on the space of Gaussian probability measures (van der Hoeven et al., 2018). In this sense our algorithm can be viewed as a modification of the kernel-based method by Bubeck et al. (2017), with the kernel estimators replaced by a quadratic approximation. This adaptation leads to enormous simplifications, both algorithmically and analytically, thanks largely to the elegant collapse of continuous exponential weights to online Newton step in the special case of unconstrained optimization with quadratic losses and Gaussian distributions.

In the algorithm and its analysis, we use the following constants (recall that C and c are suitably large/small absolute constants and $\mathrm{P} = \max(2, n, d, r)^m$ for a sufficiently large absolute constant m):

$$\mathrm{W}_{\max} = \sqrt{\frac{8d\log(4\mathrm{P})}{3}}\,, \quad \mathrm{D}_{\max} = 8\left(1 + \frac{r}{d}\right)\sqrt{\log(4\mathrm{P})}\,, \quad \eta = \frac{\mathrm{c}}{\mathrm{D}_{\max}} \min\left(\sqrt{\frac{d}{n}}, \frac{1}{d\sqrt{\log(\mathrm{P})}}\right),$$

$$\mathrm{F}_{\max} = \mathrm{C}d^2 \log(\mathrm{P})^3\,, \qquad \lambda = \frac{\mathrm{c}}{\sqrt{\mathrm{F}_{\max}\log(\mathrm{P})}}\,.$$

input n, r, $x_\circ$
let $\mu_1 = \mu_2 = x_\circ$ and $\Sigma_1 = \Sigma_2 = \frac{r^2}{d^2}\mathbb{1}$
sample X_1 from $\mathcal{N}(\mu_1, \Sigma_1)$ and observe $Y_1 = f(X_1) + \varepsilon_1$
for $t = 2$ **to** n
  sample and play X_t from $\mathcal{N}(\mu_t, \Sigma_t)$ and observe $Y_t = f(X_t) + \varepsilon_t$
  compute $W_t = \Sigma_t^{-1/2}(X_t - \mu_t)$
  compute $T_t = \boldsymbol{1}(|Y_t - Y_{t-1}| \leq \mathrm{D}_{\max} \text{ and } \|W_t\| \leq \mathrm{W}_{\max})$
  compute $D_t = T_t(Y_t - Y_{t-1})$
  $g_t = D_t \Sigma_t^{-1}(X_t - \mu_t)$ # gradient estimate
  $H_t = \lambda D_t \Sigma_t^{-1/2}\left(W_t W_t^\top - \mathbb{1}\right)\Sigma_t^{-1/2}$ # Hessian estimate
  update $\Sigma_{t+1}^{-1} = \Sigma_t^{-1} + \frac{1}{4}\eta H_t$ and $\mu_{t+1} = \mu_t - \eta\Sigma_{t+1}g_t$
end for

Algorithm 1

Let us make some remarks on the unusual features of the algorithm as well as computation and parameter choices:

(a) The algorithm uses the loss differences $Y_t - Y_{t-1}$ between consecutive rounds. This is a variance reduction trick to replace the dependence on the magnitude of the losses (on which we made no assumptions) to a dependence on the span of the losses over suitably sized balls. The latter is controlled using our assumption that the loss is Lipschitz.

(b) The loss differences and the W_t vectors are truncated if they are large, which ever so slightly biases the gradient and Hessian estimates. Algorithmically this is unnecessary as we prove the truncation occurs with negligible probability. We leave it for convenience and because it simplifies a little the analysis without impacting practical performance.

(c) The Hessian estimate H_t is symmetric but not positive definite. Despite this, the choices of η, λ and the truncation levels $\mathrm{D}_{\max}$ and $\mathrm{W}_{\max}$ ensure that Σ_t (and its inverse) remain positive definite.

(d) The computational complexity is dominated by the eigendecomposition of Σ_t, which using practical methods is $O(d^3)$ floating point operations. The space complexity is $O(d^2)$.

(e) The theoretically justified recommendations for η and λ contain universal constants that we did not explicitly calculate. The reason is that the degree of the logarithmic term is excessively conservative. We cautiously recommend dropping all the log factors and constants, which arise from (presumably) conservative high probability bounds. This would give

$$\mathrm{W}_{\max} = \infty\,, \qquad \mathrm{D}_{\max} = \infty\,, \qquad \eta = \frac{1}{1 + r/d}\sqrt{\frac{d}{n}}\,, \qquad \lambda = \frac{1}{d}\,.$$

Brief experiments suggest the algorithm remains stable with these choices but if the algorithm eventually becomes useful, then either the theory can be fine-combed to optimise the constants or better choices can be found empirically. Even better would be to find a crisper analysis that does not rely on an inductive high probability argument.

3. Surrogate loss function

We start by reintroducing the surrogate loss function used by Bubeck et al. (2017) and Lattimore and György (2021). Let $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$ be positive definite and $p = \mathcal{N}(\mu, \Sigma)$. Given $\lambda \in (0, 1)$, define

$$s(z) = \int_{\mathbb{R}^d} \left[\left(1 - \frac{1}{\lambda}\right) f(x) + \frac{1}{\lambda} f((1-\lambda)x + \lambda z)\right] p(x)\,\mathrm{d}x\,.$$

Lattimore and György (2021) noted that s is convex and $s(x) \leq f(x)$ for all x, both of which follow almost immediately from convexity of f (Figure 1). Because s is defined by a convolution with a Gaussian and f is Lipschitz, s is also infinitely differentiable. In general, s is a good approximation of f when the latter is close to linear and a poor approximation when f has considerable curvature. The next lemma collects a variety of properties of the surrogate loss (proof in Appendix A).

Lemma 3 *Suppose that Z has law $q = \mathcal{N}(\mu, \beta^2 \Sigma)$ with $\beta^2 = (2-\lambda)/\lambda$ and X has law p. Then*
(a) *s is convex and $s(z) \leq f(z)$ for all $z \in \mathbb{R}^d$;*
(b) $\mathbb{E}[s(Z)] = \mathbb{E}[f(X)]$*;*
(c) $\mathbb{E}[\nabla s(Z)] = \mathbb{E}[f(X)\Sigma^{-1}(X - \mu)]$*;*
(d) $\mathbb{E}[\nabla^2 s(Z)] = \lambda \mathbb{E}[f(X)\Sigma^{-1}((X-\mu)(X-\mu)^\top \Sigma^{-1} - \mathbb{1})]$*;*

(e) $\|\nabla^2 s(z)\| \leq \|\Sigma^{-1/2}\|$ *for all* $z \in \mathbb{R}^d$ *;*
(f) *Suppose* $z, w \in \mathbb{R}^d$ *and* $\varepsilon = \frac{\lambda}{1-\lambda}(z - w)$ *satisfies* $\|\varepsilon\|^2_{\Sigma^{-1}} \leq \frac{\log(2)^2}{2\log(P)}$*. Then*

$$\nabla^2 s(z) \leq 2\nabla^2 s(w) + \frac{\|\Sigma^{-1/2}\|}{P}\mathbb{1}\,.$$

(g) $\mathbb{E}[\langle \nabla s(Z), Z - \mu\rangle] = \beta^2 \mathbb{E}[\mathrm{tr}(\Sigma \nabla^2 s(Z))]$.

Note that if f is twice differentiable, then *(c)* equals $\mathbb{E}[\nabla f(X)]$ and *(d)* equals $\mathbb{E}[\nabla^2 f(X)]$.

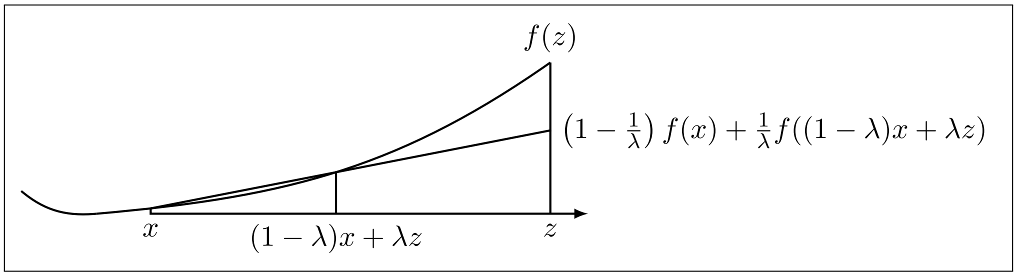

Figure 1: Given a fixed z and x, let $y = (1-\lambda)x + \lambda z$. A lower bound of $f(z)$ can be found by evaluating the second coordinate of the linear function through $(x, f(x))$ and $(y, f(y))$ at z, which is $(1 - \frac{1}{\lambda})f(x) + \frac{1}{\lambda}f(y)$. Then $s(z)$ is the average of this value over all x when x has law p.

4. Proof of Theorem 1

Assume without loss of generality that $x_\star = \mathbf{0}$, which means that the initialisation of the algorithm $x_\circ$ satisfies $\|x_\circ\| \leq r$. By construction the algorithm uses the first round to initialise the baseline and does not update the iterate. Awkwardly this means we need to bound the regret in the first round separately. Let

$$\Delta_t = \mathbb{E}_{t-1}[f(X_t) - f(x_\star)]\,,$$

which is the expected instantaneous expected regret. Since f is Lipschitz and by Jensen's inequality,

$$\Delta_1 = \mathbb{E}_{t-1}[f(X_1) - f(\mathbf{0})] \leq \mathbb{E}_{t-1}[\|X_1\|] \leq \sqrt{\mathbb{E}_{t-1}[\|X_1\|^2]} = \sqrt{\|x_\circ\|^2 + \mathrm{tr}(\Sigma_1)} \leq \sqrt{2}r\,. \quad (2)$$

Bounding the regret for the last $n-1$ rounds follows the classical analysis of mirror descent. Let $\|\cdot\|_t = \|\cdot\|_{\Sigma_t^{-1}}$. The main conceptual challenge is proving that with high probability

$$\frac{1}{2}\|\mu_{t+1}\|^2_{t+1} \leq \mathrm{F}_{\max} - \eta \sum_{s=2}^{t} \Delta_s \quad (3)$$

holds for all t, Rearranging Equation (3) with $t = n$ yields a bound on the regret. Just as important, however, is that Equation (3) ensures that the optimal point $x_\star = \mathbf{0}$ lies in the *focus region* $\{\nu \in \mathbb{R}^d : \frac{1}{2}\|\nu - \mu_t\|^2_t \leq \mathrm{F}_{\max}\}$, which is the region in which the surrogate loss function behaves more-or-less like a quadratic. Essentially we prove Equation (3) holds with high probability by induction, using in the inductive step that the optimal point lies in the focus region and hence the estimator is well-behaved.

Definition 4 *Let E_t be the event that*

(a) $\Sigma_{t+1} \leq 2\Sigma_1$ *and* $\Sigma_{t+1} \leq 2\Sigma_t$ *;*

(b) $\operatorname{tr}\left(\Sigma_{t+1}^{-1}\right) \leq \Sigma_{\max}^{-1} \triangleq \left(\frac{nd^2}{r^2} + \frac{dn^2}{4} + n\right)^2$ *;*

(c) $|\mathbb{E}_t[Y_{t+1}] - Y_t| \leq \frac{1}{2} D_{\max}$ *.*

Define a stopping time τ as the first round t where either E_t does not hold or

$$\frac{1}{2}\|\mu_{t+1}\|_{t+1}^2 \geq F_{\max} - \eta \sum_{s=1}^{t} \Delta_s \,.$$

If neither condition ever holds, then τ is defined to be n.

Note that Σ_{t+1} and μ_{t+1} are both $\mathscr{F}_t$-measurable, so τ really is a stopping time with respect to the filtration $(\mathscr{F}_t)_{t=1}^n$. Properties *(a)* and *(b)* indicate that neither Σ_{t+1} nor its inverse grows too large. Property *(c)* indicates that the losses do not change dramatically from one round to the next. In other words, properties *(a)*-*(c)* are indicators that the algorithm is stable. The following lemma, proved in Appendix D, shows that the algorithm is stable with high probability.

Lemma 5 $\mathbb{P}(\cap_{t=1}^{\tau} E_t) \geq 1 - 4/n$.

Let $\beta^2 = (2-\lambda)/\lambda$ and Z_t be a random variable that is independent of X_t and under $\mathbb{P}_{t-1}$ has law $\mathcal{N}(\mu_t, \beta^2 \Sigma_t)$, and define the surrogate loss at time t as

$$s_t(z) = \mathbb{E}_{t-1}\left[\left(1 - \frac{1}{\lambda}\right) f(X_t) + \frac{1}{\lambda} f((1-\lambda)X_t + \lambda z)\right] \,.$$

The truncation in the gradient and Hessian estimators introduces a small amount of bias that needs to be controlled (proof in Appendix B).

Lemma 6 *On $\{2 \leq t \leq \tau\}$ and for a positive definite matrix A,*

(a) $\Big|\mathbb{E}_{t-1}\left[\langle g_t, \mu_t\rangle\right] - \mathbb{E}_{t-1}\left[\langle \nabla s_t(Z_t), \mu_t\rangle\right]\Big| \leq \dfrac{\lambda}{n}$ *;*

(b) $\Big|\mathbb{E}_{t-1}\left[\operatorname{tr}(A H_t)\right] - \mathbb{E}_{t-1}\left[\operatorname{tr}(A \nabla^2 s_t(Z_t))\right]\Big| \leq \operatorname{tr}(A\Sigma_t^{-1}) \min\left(\dfrac{\lambda}{2nF_{\max}}, \dfrac{\lambda}{dn}, \dfrac{1}{n\Sigma_{\max}^{-1}}\right)$.

Step 1: High-level argument Expanding the square shows that

$$\frac{1}{2}\|\mu_{t+1}\|_{t+1}^2 - \frac{1}{2}\|\mu_t\|_t^2 = \underbrace{-\eta\langle g_t, \mu_t\rangle + \frac{\eta}{8}\|\mu_t\|_{H_t}^2}_{A_t} + \underbrace{\frac{\eta^2}{2}\|g_t\|_{\Sigma_{t+1}}^2}_{B_t} \,. \tag{4}$$

A_t collects those terms that are linear in the learning rate and B_t those that are quadratic. The expectation of the linear term will be shown to be close to $-\eta\Delta_t$ (recall that $\Delta_t = \mathbb{E}_{t-1}[f(X)] - f(\mathbf{0})$ is the expected instantaneous regret). The lower order terms will be shown to be $O(\eta^2)$. Besides technical complications, the result follows by dividing both sides by the learning rate, rearranging and telescoping the potentials. The principle difficulty is that our bounds on A_t and B_t only hold when $\frac{1}{2}\|\mu_t\|_t^2$ is not too large, which has to be tracked through the analysis with induction and a high probability argument.

Step 2: Linear terms This step is the most fundamental. We show that the linear terms can be controlled in terms of the regret and some small correction terms.

Lemma 7 *With probability at least* $1 - 1/n$*,*

$$\sum_{t=2}^{\tau} A_t \leq -\eta \sum_{t=2}^{\tau} \Delta_t + \eta\beta^2 \sum_{t=2}^{\tau} \operatorname{tr}(\Sigma_t H_t) + 1100\eta D_{\max} F_{\max}^{1/2} \sqrt{n \log(P)} \,.$$

Proof Suppose that $2 \leq t \leq \tau$. Remember that Z_t is a random element that under $\mathbb{P}_{t-1}$ has law $\mathcal{N}(\mu_t, \beta^2 \Sigma_t)$ and is independent from X_t. Then

$$\begin{aligned}
\mathbb{E}_{t-1}[A_t] &= \mathbb{E}_{t-1}\left[-\eta\langle g_t, \mu_t\rangle + \frac{\eta}{8}\|\mu_t\|^2_{H_t}\right] \\
&\leq \mathbb{E}_{t-1}\left[-\eta\langle \nabla s_t(Z_t), \mu_t\rangle + \frac{\eta}{8}\|\mu_t\|^2_{\nabla^2 s_t(Z_t)}\right] + \frac{2\eta\lambda}{n} && \text{Lemma 6ab} \\
&= \mathbb{E}_{t-1}\left[-\eta\langle \nabla s_t(Z_t), Z_t\rangle + \eta\langle \nabla s_t(Z_t), Z_t - \mu_t\rangle + \frac{\eta}{8}\|\mu_t\|^2_{\nabla^2 s_t(Z_t)}\right] + \frac{2\eta\lambda}{n} \\
&= \mathbb{E}_{t-1}\left[-\eta\langle \nabla s_t(Z_t), Z_t\rangle + \eta\beta^2 \operatorname{tr}(\Sigma_t \nabla^2 s_t(Z_t)) + \frac{\eta}{8}\|\mu_t\|^2_{\nabla^2 s_t(Z_t)}\right] + \frac{2\eta\lambda}{n} && \text{Theorem 3g} \\
&\leq \mathbb{E}_{t-1}\left[-\eta\langle \nabla s_t(Z_t), Z_t\rangle + \eta\beta^2 \operatorname{tr}(\Sigma_t H_t) + \frac{\eta}{8}\|\mu_t\|^2_{\nabla^2 s_t(Z_t)}\right] + \frac{4\eta}{n}\,, && \text{Lemma 6b}
\end{aligned}$$

where in the first inequality we used that $\operatorname{tr}(\mu\mu^\top \Sigma_t^{-1}) \leq 2\mathrm{F}_{\max}$ since $t \leq \tau$ and in the second we used $\lambda \leq 1$ and $\lambda\beta^2 = 2 - \lambda \leq 2$. To make progress on bounding the first term, recall that s_t is infinitely differentiable. Hence, by Taylor's theorem, for all $z \in \mathbb{R}^d$ there exists a $\xi_z \in [\mathbf{0}, z] = \{\alpha z : \alpha \in [0, 1]\}$ such that

$$s_t(\mathbf{0}) = s_t(z) - \langle \nabla s_t(z), z\rangle + \frac{1}{2}\|z\|^2_{\nabla^2 s_t(\xi_z)} \,.$$

We need a simple lemma (proof in Appendix C based on Lemma 3f) to bound $\nabla^2 s_t(\xi_z)$.

Lemma 8 *On* $\{2 \leq t \leq \tau\}$*,*

(a) $\mathbb{E}_{t-1}\left[\|Z_t\|^2_{\nabla^2 s_t(\xi_{Z_t})}\right] \geq \frac{1}{2}\mathbb{E}_{t-1}\left[\|Z_t\|^2_{\nabla^2 s_t(\mu_t)}\right] - \frac{1}{n}$.

(b) $\mathbb{E}_{t-1}\left[\|\mu_t\|^2_{\nabla^2 s_t(\mu_t)}\right] \geq \frac{1}{2}\mathbb{E}_{t-1}\left[\|\mu_t\|^2_{\nabla^2 s_t(Z_t)}\right] - \frac{1}{n}$.

Using Lemma 8,

$$\begin{aligned}
\mathbb{E}_{t-1}[\langle \nabla s_t(Z_t), Z_t\rangle] &= \mathbb{E}_{t-1}\left[s_t(Z_t) - s_t(\mathbf{0}) + \frac{1}{2}\|Z_t\|^2_{\nabla^2 s_t(\xi_{Z_t})}\right] \\
&\geq \mathbb{E}_{t-1}\left[s_t(Z_t) - s_t(\mathbf{0}) + \frac{1}{4}\|Z_t\|^2_{\nabla^2 s_t(\mu_t)}\right] - \frac{1}{n} && \text{Lemma 8a} \\
&= \mathbb{E}_{t-1}\left[s_t(Z_t) - s_t(\mathbf{0}) + \frac{\beta^2}{4}\operatorname{tr}(\Sigma_t \nabla^2 s_t(\mu_t)) + \frac{1}{4}\|\mu_t\|^2_{\nabla^2 s_t(\mu_t)}\right] - \frac{1}{n} \\
&\geq \mathbb{E}_{t-1}\left[s_t(Z_t) - s_t(\mathbf{0}) + \frac{1}{8}\|\mu_t\|^2_{\nabla^2 s_t(Z_t)}\right] - \frac{2}{n} && \text{Lemma 8b} \\
&\geq \mathbb{E}_{t-1}\left[f(X_t) - f(\mathbf{0}) + \frac{1}{8}\|\mu_t\|^2_{\nabla^2 s_t(Z_t)}\right] - \frac{2}{n} && \text{Lemma 3ab} \\
&= \Delta_t + \frac{1}{8}\mathbb{E}_{t-1}\left[\|\mu_t\|^2_{\nabla^2 s_t(Z_t)}\right] - \frac{2}{n}\,, && \text{Definition of } \Delta_t
\end{aligned}$$

where in the second equality we used the fact that Z_t has law $\mathcal{N}(\mu_t, \beta^2\Sigma_t)$ under $\mathbb{P}_{t-1}$, and in the second inequality that the matrices in the dropped term are positive semi-definite. Hence,

$$\mathbb{E}_{t-1}[A_t] \leq -\eta\Delta_t + \eta\beta^2\mathbb{E}_{t-1}[\operatorname{tr}(\Sigma_t H_t)] + \frac{6\eta}{n}\,.$$

This shows the connection between the linear component of the change in the potential and the regret. The remainder of the proof of the lemma is devoted to a concentration analysis converting the bound in expectation to something that holds with high probability. By the above display and the definition of A_t,

$$\begin{aligned}
\sum_{t=2}^{\tau} A_t \leq \sum_{t=2}^{\tau}[A_t - \mathbb{E}_{t-1}[A_t]] + \eta\beta^2\sum_{t=2}^{\tau}\left(\mathbb{E}_{t-1}[\operatorname{tr}(\Sigma_t H_t)] - \operatorname{tr}(\Sigma_t H_t)\right) \\
+ \eta\beta^2\sum_{t=2}^{\tau}\operatorname{tr}(\Sigma_t H_t) - \eta\sum_{t=2}^{\tau}\Delta_t + 6\eta\,.
\end{aligned}$$

The first two terms on the right-hand side are sums of martingale differences, which we now control using concentration of measure. We need to show that the tails of A_t and $\operatorname{tr}(\Sigma_t H_t)$ are well-behaved under $\mathbb{P}_{t-1}$ whenever $t \leq \tau$. Assume that $t \leq \tau$. Then, since $D_t \leq \mathrm{D}_{\max}$ and $\frac{1}{2}\|\mu_t\|_t^2 \leq \mathrm{F}_{\max}$, Fact 11a implies

$$\|\eta\langle \mu_t, g_t\rangle\|_{t-1,\psi_2} = \|\eta D_t\langle \Sigma_t^{-1/2}\mu_t, W_t\rangle\|_{t-1,\psi_2} \leq 2\eta\mathrm{D}_{\max}\|\mu_t\|_t \leq 3\eta\mathrm{D}_{\max}\mathrm{F}_{\max}^{1/2}\,.$$

Lemma 2.7.7 in the book by Vershynin (2018) says that $\|XY\|_{\psi_1} \leq \|X\|_{\psi_2}\|Y\|_{\psi_2}$ for any random variables X and Y. By definition $\|1\|_{\psi_2} = 1/\sqrt{\log(2)}$. Combining these with the above display shows that

$$\|\eta\langle \mu_t, g_t\rangle\|_{t-1,\psi_1} \leq 4\eta\mathrm{D}_{\max}\mathrm{F}_{\max}^{1/2}\,.$$

Next, using Fact 11b, $\|1\|_{\psi_1} = 1/\log(2)$, and that $\lambda \leq \mathrm{F}_{\max}^{-1/2}$,

$$\left\|\frac{\eta}{8}\|\mu_t\|^2_{H_t}\right\|_{t-1,\psi_1} = \left\|\frac{\eta\lambda D_t}{8}\|\Sigma_t^{-1/2}\mu_t\|^2_{W_tW_t^\top - \mathbb{1}}\right\|_{t-1,\psi_1} \leq \eta\lambda\mathrm{D}_{\max}\|\mu_t\|_t^2 \leq 2\eta\mathrm{D}_{\max}\mathrm{F}_{\max}^{1/2}\,.$$

Combining the above two displays with the triangle inequality implies that $\|A_t\|_{t-1,\psi_1} \leq 6\eta \mathrm{D}_{\max} \mathrm{F}_{\max}^{1/2}$. Lastly, since $\lambda\beta^2 = 2 - \lambda \leq 2$ and $d \leq \mathrm{F}_{\max}^{1/2}$,

$$\left\|\eta\beta^2 \operatorname{tr}(\Sigma_t H_t)\right\|_{t-1,\psi_1} = \eta\lambda\beta^2 \left\|D_t \operatorname{tr}(W_t W_t^\top - \mathbb{1})\right\|_{t-1,\psi_1} \leq 5\eta\lambda\beta^2 d \mathrm{D}_{\max} \leq 10\eta \mathrm{D}_{\max} \mathrm{F}_{\max}^{1/2}.$$

The claim of the lemma now follows by Bernstein's inequality (Lemma 13) and naively simplifying the constants. ■

Step 3: Quadratic terms The next step is to bound the quadratic terms in Equation 4.

Lemma 9 $\sum_{t=2}^{\tau} B_t \leq n\eta^2 D_{\max}^2 W_{\max}^2$.

Proof Suppose $2 \leq t \leq \tau$. Then,

$$B_t = \frac{\eta^2}{2}\|g_t\|_{\Sigma_{t+1}}^2 = \frac{\eta^2 D_t^2}{2}\|\Sigma_t^{-1/2}(X_t - \mu_t)\|_{\Sigma_t^{-1/2}\Sigma_{t+1}\Sigma_t^{-1/2}}^2 \leq \eta^2 \mathrm{D}_{\max}^2 \mathrm{W}_{\max}^2.$$

where in the final inequality we used that on $\{t \leq \tau\}$, $\Sigma_{t+1} \leq 2\Sigma_t$ and $D_t\|W_t\| \leq \mathrm{D}_{\max}\mathrm{W}_{\max}$. The result follows because $\tau \leq n$ by definition. ■

Step 4: Bounding the regret By Lemma 5, Lemma 7 and Lemma 9, with probability least $1 - 5/n$, $\bigcap_{t=1}^{\tau} E_t$ holds and

$$\begin{aligned}
\sum_{t=2}^{\tau} \frac{1}{2}\|\mu_{t+1}\|_{t+1}^2 - \frac{1}{2}\|\mu_t\|_t^2 &= \sum_{t=2}^{\tau}(A_t + B_t) \\
&\leq 1100\eta \mathrm{D}_{\max}\mathrm{F}_{\max}^{1/2}\sqrt{n\log(\mathrm{P})} + n\eta^2 \mathrm{D}_{\max}^2 \mathrm{W}_{\max}^2 + \eta\beta^2 \sum_{t=2}^{\tau} \operatorname{tr}(\Sigma_t H_t) - \eta \sum_{t=2}^{\tau} \Delta_t \\
&\leq 1100c\mathrm{F}_{\max}^{1/2}\sqrt{d\log(\mathrm{P})} + c^2 d\mathrm{W}_{\max}^2 + \eta\beta^2 \sum_{t=2}^{\tau} \operatorname{tr}(\Sigma_t H_t) - \eta \sum_{t=2}^{\tau} \Delta_t\,,
\end{aligned}$$

where the second equality follows from the definition of η. Using that $x \leq 2\log(1+x)$ for $x \in [0,1]$ it follows that for any positive definite matrix X with $\|X\| \leq 1$, $\operatorname{tr}(X) \leq 2\log\det(\mathbb{1} + X)$. Therefore, on the event $\cap_{t=1}^{\tau} E_t$,

$$\begin{aligned}
\eta\beta^2 \sum_{t=2}^{\tau} \operatorname{tr}(\Sigma_t H_t) &\leq 8\beta^2 \sum_{t=2}^{\tau} \log\det\left(\mathbb{1} + \frac{\eta}{4}\Sigma_t H_t\right) \\
&= 8\beta^2 \log\left(\det\left(\Sigma_1 \Sigma_{\tau+1}^{-1}\right)\right) \\
&\leq \frac{16d}{\lambda}\log\left(\frac{\operatorname{tr}\left(\Sigma_1\Sigma_{\tau+1}^{-1}\right)}{d}\right) && \text{Jensen's inequality} \\
&\leq \frac{16d\mathrm{F}_{\max}^{1/2}\log(\mathrm{P})^{3/2}}{c}\,,
\end{aligned}$$

where the equality holds because $\Sigma_t \Sigma_{t+1}^{-1} = \mathbb{1} + \frac{\eta}{4} \Sigma_t H_t$ by the update for Σ_{t+1}^{-1}, in the second inequality we used $\beta^2 = (2-\lambda)/\lambda \leq 2/\lambda$, and the last inequality holds because $\operatorname{tr}(\Sigma_{\tau+1}^{-1}) \leq \Sigma_{\max}^{-1}$ by Definition 4b, the definition of Σ_1, and by choosing P large enough. Therefore with probability at least $1 - 5/n$, E_τ holds and

$$\begin{aligned}\frac{1}{2}\|\mu_{\tau+1}\|_{\tau+1}^2 &\leq \frac{1}{2}\|\mu_1\|_1^2 + \frac{16d\mathrm{F}_{\max}^{1/2}}{c}\log(\mathrm{P})^{3/2} + 1100c\mathrm{F}_{\max}^{1/2}\sqrt{d\log(\mathrm{P})} + c^2 d\mathrm{W}_{\max}^2 - \eta\sum_{t=2}^{\tau}\Delta_t \\ &< \mathrm{F}_{\max} - \eta\sum_{t=2}^{\tau}\Delta_t\,,\end{aligned}$$

where in the second inequality we used the definition of $\mathrm{F}_{\max}$ and by choosing C suitably large and the fact that $\frac{1}{2}\|\mu_1\|_1^2 \leq d^2/2$. Since $\{\frac{1}{2}\|\mu_{\tau+1}\|_{\tau+1}^2 \leq \mathrm{F}_{\max} - \eta\sum_{t=2}^{\tau}\Delta_t\} \cap E_\tau$ implies that $\tau = n$, it follows by rearranging the above display that

$$\mathbb{P}\left(\sum_{t=2}^{n}\Delta_t \leq \frac{\mathrm{F}_{\max}}{\eta} \quad \text{and} \quad \tau = n\right) \geq 1 - \frac{5}{n}\,. \tag{5}$$

The last step is to bound the actual regret in terms $\sum_{t=1}^{n}\Delta_t$. By Lemma 15a, on $\{2 \leq t \leq \tau\}$,

$$\|f(X_t) - \mathbb{E}_{t-1}[f(X_t)]\|_{t-1,\psi_2} \leq 2\|\Sigma_t\|^{1/2} \leq \frac{3r}{d}\,.$$

Therefore by Lemma 13, with probability at least $1 - 1/n$,

$$\sum_{t=1}^{\tau} f(X_t) - f(\mathbf{0}) \leq \sum_{t=1}^{\tau}\Delta_t + \frac{200r}{d}\sqrt{n\log(n)}\,.$$

Combining the above with Equation (2) and Equation (5) implies that with probability at least $1 - 6/n$,

$$\sum_{t=1}^{n} f(X_t) - f(\mathbf{0}) \leq \frac{\mathrm{F}_{\max}}{\eta} + \frac{200r}{d}\sqrt{n\log(n)} + \sqrt{2}r\,.$$

Theorem 1 now follows from the definitions of η and $\mathrm{F}_{\max}$.

5. Discussion

There are a few outstanding issues.

Handling constraints Our algorithm cannot handle constraints on the domain of f. The principle problem is that the algorithm samples its actions from a Gaussian distribution, and when the losses have low curvature the covariance of this Gaussian could be large enough that the algorithm plays outside the constraint set with some non-negligible probability. There are several ideas one may try. For example, by estimating some kind of extension of f or regularising to prevent the focus region from leaving the domain. It would surprise us if nothing can be made to work, possibly at the price of a worse dimension-dependence.

Adversarial setting Algorithms based on elimination or focus regions cannot handle the adversarial setting without some sort of correction. Bubeck et al. (2017) and Suggala et al. (2021) both use restarts, which may also be usable in our setting. Note that in the adversarial version of the problem the centering of the gradient/Hessian estimators using the loss from the previous round no longer makes sense and the dependence on d in front of the diameter should be expected to increase slightly.

Dependence on various quantities A natural question is whether or not there is scope to improve the bound. With these techniques, it feels like there is limited room for improvement. In particular the bounds on the stability and variance of the algorithm seem to be tight. There is *still* no lower bound that is superlinear in the dimension. Maybe the true dimension dependence is linear in d, but fundamentally new ideas seem to be needed for such a result. One may also wonder about the dependence on r. The quantity r/d is effectively the range of the observed losses. Because our setting is unconstrained, we cannot assume the losses are globally bounded in $[0, 1]$ as is standard in the constrained setting. Our expectation is that once the analysis is applied to the constrained case, the quantity r/d will be replaced by 1. In fact, we believe that even in the current setting the r/d term can be moved to a lower-order term by exploiting the fact that eventually the loss function must be very flat on the focus region, since otherwise the regret would not be sublinear.

Sample complexity Theorem 2 shows that $\frac{1}{n}\sum_{t=1}^n X_t$ is near-optimal with high probability for suitably large n. Using convexity one can easily show that $\frac{1}{n}\sum_{t=1}^n \mu_t$ is also near-optimal with the same sample complexity and is unsurprisingly empirically superior.

Acknowledgments

Thanks to Tim van Erven, Wouter Koolen, Dirk van der Hoeven, Jack Mayo and Hidde Fokkema for pointing out that our analysis could be simplified by using Σ_{t+1} in the update rather than Σ_t that appeared in an earlier version.

References

A. Agarwal, O. Dekel, and L. Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Colt*, pages 28–40. Citeseer, 2010.

A. Agarwal, D. P. Foster, D. Hsu, S. M. Kakade, and A. Rakhlin. Stochastic convex optimization with bandit feedback. *SIAM Journal on Optimization*, 23(1):213–240, 2013.

A. Belloni, T. Liang, H. Narayanan, and A. Rakhlin. Escaping the local minima via simulated annealing: Optimization of approximately convex functions. In *Conference on Learning Theory*, pages 240–265, 2015.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. OUP Oxford, 2013.

S. Bubeck and R. Eldan. Exploratory distributions for convex functions. *Mathematical Statistics and Learning*, 1(1):73–100, 2018.

S. Bubeck, O. Dekel, T. Koren, and Y. Peres. Bandit convex optimization: $\sqrt{T}$ regret in one dimension. In *Proceedings of the 28th Conference on Learning Theory*, pages 266–278, Paris, France, 2015. JMLR.org.

S. Bubeck, Y.T. Lee, and R. Eldan. Kernel-based methods for bandit convex optimization. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, pages 72–85, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4528-6.

V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Conference on Learning Theory*, pages 355–366, 2008.

D.C. Dowson and B.V. Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

A Flaxman, A Kalai, and HB McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *SODA'05: Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394, 2005.

E. Hazan and K. Levy. Bandit convex optimization: Towards tight bounds. In *Advances in Neural Information Processing Systems*, pages 784–792, 2014.

E. Hazan and Y. Li. An optimal algorithm for bandit convex optimization. *arXiv preprint arXiv:1603.04350*, 2016.

E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

X. Hu, Prashanth L.A., A. György, and Cs. Szepesvári. (Bandit) convex optimization with biased noisy gradient oracles. In *AISTATS*, pages 819–828, 2016.

S. Ito. An optimal algorithm for bandit convex optimization with strongly-convex and smooth loss. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2229–2239. PMLR, 26–28 Aug 2020.

R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, pages 697–704. MIT Press, 2005.

T. Lattimore. Improved regret for zeroth-order adversarial bandit convex optimisation. *Mathematical Statistics and Learning*, 2(3/4):311–334, 2020.

T. Lattimore and A. György. Improved regret for zeroth-order stochastic convex bandits. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2938–2964. PMLR, 15–19 Aug 2021.

H. Luo, M. Zhang, and P. Zhao. Adaptive bandit convex optimization with heterogeneous curvature. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1576–1612. PMLR, 02–05 Jul 2022.

D. Russo and B. Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, pages 1583–1591. Curran Associates, Inc., 2014.

A. Saha and A. Tewari. Improved regret guarantees for online smooth convex optimization with bandit feedback. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 636–642, 2011.

A. Suggala, P. Ravikumar, and P. Netrapalli. Efficient bandit convex optimization: Beyond linear losses. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4008–4067. PMLR, 15–19 Aug 2021.

D. van der Hoeven, T. van Erven, and W. Kotłowski. The many faces of exponential weights in online learning. In *Proceedings of the 31st Conference on Learning Theory*, pages 2067–2092, 2018.

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

Appendix A. Proof of Lemma 3

The proof is complicated slightly because we have not assumed that f is differentiable. For $\delta > 0$ let $p_\delta = \mathcal{N}(\mathbf{0}, \delta \mathbb{1})$ and $f_\delta(x) = \int_{\mathbb{R}^d} f(x+y) p_\delta(y) \, \mathrm{d}y$ be the convolution of f and the Gaussian p_δ. Note that f_δ is infinitely differentiable and inherits convexity and Lipshitzness from f (all immediate from definitions and a good exercise). Further, f_δ converges uniformly to $f_0 \triangleq f$ as $\delta \to 0$. For any $\delta \geq 0$, let

$$s_\delta(z) = \mathbb{E}\left[\left(1 - \frac{1}{\lambda}\right) f_\delta(X) + \frac{1}{\lambda} f_\delta((1-\lambda)X + \lambda z)\right] ,$$

which is the surrogate loss function associated with f_δ. By a change of variable, $u = x + \frac{\lambda}{1-\lambda} z$,

$$s_\delta(z) = \frac{1}{\lambda} \int_{\mathbb{R}^d} f_\delta((1-\lambda)x + \lambda z) p(x) \, \mathrm{d}x = \frac{1}{\lambda} \int_{\mathbb{R}^d} f_\delta((1-\lambda)u) p\left(u - \frac{\lambda}{1-\lambda} z\right) \mathrm{d}u \, .$$

Therefore, by exchanging derivatives and integrals and reversing the change of measure,

$$\begin{aligned} \nabla s_\delta(z) &= \frac{1}{1-\lambda} \int_{\mathbb{R}^d} f_\delta((1-\lambda)u) \Sigma^{-1} \left(u - \frac{\lambda}{1-\lambda} z - \mu\right) p\left(u - \frac{\lambda}{1-\lambda} z\right) \mathrm{d}u \\ &= \frac{1}{1-\lambda} \int_{\mathbb{R}^d} f_\delta((1-\lambda)x + \lambda z) \Sigma^{-1}(x - \mu) p(x) \, \mathrm{d}x \, . \end{aligned}$$

Using the uniform convergence of f_δ to f as $\delta \to 0$ yields $\lim_{\delta \to 0} \|\nabla s_\delta(z) - \nabla s(z)\| = 0$ for all $z \in \mathbb{R}^d$. For the Hessian,

$$\nabla^2 s_\delta(z) = \frac{\lambda}{(1-\lambda)^2} \mathbb{E}\left[f_\delta((1-\lambda)X + \lambda z) \Sigma^{-1}((X-\mu)(X-\mu)^\top \Sigma^{-1} - \mathbb{1})\right] .$$

Hence, $\lim_{\delta \to 0} \|\nabla^2 s_\delta(z) - \nabla^2 s(z)\| = 0$.

(a) Convexity of s and $s(z) \leq f(z)$ for all $z \in \mathbb{R}^d$ follow from the definition of s and the convexity of f, as was noted already by Lattimore and György (2021), with the intuition given in Figure 1.

(b) By definition $(1-\lambda)X + \lambda Z$ has the same law as X. Therefore,

$$\mathbb{E}[s(Z)] = \mathbb{E}\left[\left(1 - \frac{1}{\lambda}\right) f(X) + \frac{1}{\lambda} f((1-\lambda)X + \lambda Z)\right] = \mathbb{E}\left[f(X)\right] .$$

(c) By exchanging the integral and derivative, for any $\delta > 0$,

$$\mathbb{E}\left[\nabla s_\delta(Z)\right] = \mathbb{E}\left[\nabla f_\delta((1-\lambda)X + \lambda Z)\right] = \mathbb{E}\left[\nabla f_\delta(X)\right] = \mathbb{E}\left[f_\delta(X)\Sigma^{-1}(X - \mu)\right] .$$

Taking the limit as $\delta \to 0$ establishes the part.[1]

(d) As above,

$$\begin{aligned}
\mathbb{E}\left[\nabla^2 s_\delta(Z)\right] &= \lambda \mathbb{E}\left[\nabla^2 f_\delta((1-\lambda)X + \lambda Z)\right] \\
&= \lambda \mathbb{E}\left[\nabla^2 f_\delta(X)\right] \\
&= \lambda \mathbb{E}\left[f_\delta(X)\Sigma^{-1}((X-\mu)(X-\mu)^\top \Sigma^{-1} - \mathbb{1})\right] .
\end{aligned}$$

Taking the limit again completes the part.

(e) This is a consequence of the Lipschitzness of f: Let $u \in \mathbb{R}^d$ have $\|u\| = 1$. Since f is Lipschitz, so is f_δ, which means that $\|\nabla f_\delta(w)\| \leq 1$ and $|\langle u, \nabla f_\delta(w)\rangle| \leq 1$ for any $w \in \mathbb{R}^d$. Therefore, for any $z \in \mathbb{R}^d$,

$$\begin{aligned}
u^\top \nabla^2 s_\delta(z) u &= \frac{\lambda}{1-\lambda} \mathbb{E}\left[\langle u, \nabla f_\delta((1-\lambda)X + \lambda z)\rangle \langle u, \Sigma^{-1}(X-\mu)\rangle\right] \\
&\leq \frac{\lambda}{1-\lambda} \mathbb{E}\left[\left|\langle u, \Sigma^{-1}(X-\mu)\rangle\right|\right] \\
&\leq \frac{\lambda}{1-\lambda} \sqrt{\mathbb{E}\left[\langle u, \Sigma^{-1}(X-\mu)\rangle^2\right]} \\
&= \frac{\lambda}{1-\lambda} \sqrt{\|u\|^2_{\Sigma^{-1}}} \\
&\leq \frac{\lambda}{1-\lambda} \|\Sigma^{-1/2}\| .
\end{aligned}$$

Therefore $\|\nabla^2 s_\delta(z)\| \leq \frac{\lambda}{1-\lambda}\|\Sigma^{-1/2}\| \leq \|\Sigma^{-1/2}\|$ for all δ and the result follows again by taking the limit as $\delta \to 0$.

(f) Recall that $\varepsilon = \frac{\lambda}{1-\lambda}(z - w)$ and define the event $E = \{x \in \mathbb{R}^d : \langle x - \mu, \Sigma^{-1}\varepsilon\rangle \leq \log(2)\}$. Then,

$$\begin{aligned}
\nabla^2 s_\delta(z) &= \lambda \int_{\mathbb{R}^d} \nabla^2 f_\delta((1-\lambda)x + \lambda z) p(x) \, \mathrm{d}x \\
&= \lambda \int_{\mathbb{R}^d} \nabla^2 f_\delta((1-\lambda)x + \lambda w) p(x - \varepsilon) \, \mathrm{d}x \\
&= \underbrace{\lambda \int_E \nabla^2 f_\delta((1-\lambda)x + \lambda w) \frac{p(x-\varepsilon)}{p(x)} p(x) \, \mathrm{d}x}_{\text{A}} + \underbrace{\lambda \int_{E^c} \nabla^2 f_\delta((1-\lambda)x + \lambda w) p(x - \varepsilon) \, \mathrm{d}x}_{\text{B}} .
\end{aligned}$$

1. Taking the limit with respect to δ is needed because unlike f_δ for $\delta > 0$, f may not be differentiable.

The first term is upper bounded as

$$\begin{aligned}
\mathrm{A} &= \lambda \int_E \nabla^2 f_\delta((1-\lambda)x + \lambda w) \frac{p(x-\varepsilon)}{p(x)} p(x)\,\mathrm{d}x \\
&= \lambda \int_E \nabla^2 f_\delta((1-\lambda)x + \lambda w) \exp\left(-\frac{1}{2}\|\varepsilon\|^2_{\Sigma^{-1}} + \langle x-\mu, \Sigma^{-1}\varepsilon\rangle\right) p(x)\,\mathrm{d}x \\
&\le 2\lambda \int_E \nabla^2 f_\delta((1-\lambda)x + \lambda w) p(x)\,\mathrm{d}x && \text{Definition of } E \\
&\le 2\lambda \int_{\mathbb{R}^d} \nabla^2 f_\delta((1-\lambda)x + \lambda w) p(x)\,\mathrm{d}x && \text{Convexity of } f \\
&= 2\nabla^2 s_\delta(w)\,.
\end{aligned}$$

To bound B, similarly to the calculations in part *(e)*, we have

$$\begin{aligned}
\|\mathrm{B}\| &= \lambda \sup_{u:\|u\|\le 1} \operatorname{tr}\left(uu^\top \int_{E^c} \nabla^2 f_\delta((1-\lambda)x+\lambda w) p(x-\varepsilon)\,\mathrm{d}x\right) \\
&= \frac{\lambda}{1-\lambda} \sup_{u:\|u\|\le 1} \int_{E^c} \langle u, \nabla f_\delta((1-\lambda)x+\lambda w)\rangle \langle u, \Sigma^{-1}(x-\mu-\varepsilon)\rangle p(x-\varepsilon)\,\mathrm{d}x \\
&\le \frac{\lambda}{1-\lambda} \sup_{u:\|u\|\le 1} \int_{E^c} \left|\langle u, \Sigma^{-1}(x-\mu-\varepsilon)\rangle\right| p(x-\varepsilon)\,\mathrm{d}x \\
&\le \frac{\lambda}{1-\lambda} \sup_{u:\|u\|\le 1} \sqrt{\int_{E^c} p(x-\varepsilon)\,\mathrm{d}x \cdot \int_{\mathbb{R}^d} \langle u, \Sigma^{-1}(x-\mu-\varepsilon)\rangle^2 p(x-\varepsilon)\,\mathrm{d}x} \\
&= \frac{\lambda}{1-\lambda} \sup_{u:\|u\|\le 1} \sqrt{\int_{E^c} p(x-\varepsilon)\,\mathrm{d}x \cdot \|u\|^2_{\Sigma^{-1}}} \\
&\le \frac{\lambda}{1-\lambda} \|\Sigma^{-1/2}\| \sqrt{\int_{E^c} p(x-\varepsilon)\,\mathrm{d}x} \\
&= \frac{\lambda}{1-\lambda} \|\Sigma^{-1/2}\| \sqrt{\mathbb{P}\left(\langle X+\varepsilon-\mu, \Sigma^{-1}\varepsilon\rangle \ge \log(2)\right)}\,,
\end{aligned}$$

where the second inequality holds by Cauchy-Schwartz, and the non-negativity of the terms in the second integral. Note that $\langle X-\mu, \Sigma^{-1}\varepsilon\rangle$ has law $\mathcal{N}(0, \|\varepsilon\|^2_{\Sigma^{-1}})$. Hence, by standard Gaussian concentration (Boucheron et al., 2013, §2.2), if

$$\mathbb{P}\left(\langle X-\mu, \Sigma^{-1}\varepsilon\rangle + \|\varepsilon\|^2_{\Sigma^{-1}} \ge \log(2)\right) \le \exp\left(-\frac{\left(\log(2) - \|\varepsilon\|^2_{\Sigma^{-1}}\right)^2}{2\|\varepsilon\|^2_{\Sigma^{-1}}}\right) \le \frac{1}{\mathrm{P}}\,,$$

where in the final inequality we used the assumption that $\|\varepsilon\|^2_{\Sigma^{-1}} \le \frac{\log(2)^2}{2\log(\mathrm{P})}$, which also implies $\|\varepsilon\|^2_{\Sigma^{-1}} \le \log(2)$ as $\mathrm{P} \ge 2$, which is necessary for the application of the concentration inequality. The result follows since $\lambda/(1-\lambda) \le 1$.

(g) No particular properties of s are needed here beyond twice differentiability and that s is Lipschitz, which ensures that $\lim_{t\to\infty} \int_{\mathbb{R}^d:\|z\|\ge t} s(z)q(z)\,\mathrm{d}z = 0$. By definition and integrating by

parts,

$$\begin{aligned}\mathbb{E}[\langle \nabla s(Z), Z-\mu\rangle] &= \operatorname{tr}\left(\int_{\mathbb{R}^d} \nabla s(z)(z-\mu)^\top q(z)\,\mathrm{d}z\right) \\ &= -\beta^2 \operatorname{tr}\left(\Sigma \int_{\mathbb{R}^d} \nabla s(z) \nabla q(z)\,\mathrm{d}z\right) \\ &= \beta^2 \operatorname{tr}\left(\Sigma \int_{\mathbb{R}^d} \nabla^2 s(z) q(z)\,\mathrm{d}z\right) \\ &= \beta^2 \mathbb{E}\left[\operatorname{tr}(\Sigma \nabla^2 s(Z))\right] ,\end{aligned}$$

where in the second equality we used the fact that $\nabla q(z) = -\beta^{-2}\Sigma^{-1}(z-\mu)q(z)$ and the cyclic property of the trace. The third equality follows using integrating by parts.

Appendix B. Proof of Lemma 6

The conceptual part of this proof is straightforward and important for understanding the main ideas. Sadly there is also a tedious part, which involves handling the truncation used in the gradient and Hessian estimates.

Conceptual part Let $I_t = 1 - T_t$. By definition,

$$\begin{aligned}\mathbb{E}_{t-1}[g_t] &= \mathbb{E}_{t-1}[D_t \Sigma_t^{-1}(X_t - \mu_t)] \\ &= \mathbb{E}_{t-1}[(Y_t - Y_{t-1})\Sigma_t^{-1}(X_t-\mu_t)] - \underbrace{\mathbb{E}_{t-1}[(Y_t - Y_{t-1}) I_t \Sigma_t^{-1}(X_t - \mu_t)]}_{\mathcal{E}_1}\end{aligned}$$

The second (error) term is intuitively small because $I_t = 0$ with overwhelming probability. Carefully bounding this is the tedious part. The first term satisfies

$$\begin{aligned}\mathbb{E}_{t-1}[(Y_t - Y_{t-1})\Sigma_t^{-1}(X_t - \mu_t)] &= \mathbb{E}_{t-1}[Y_t \Sigma_t^{-1}(X_t-\mu_t)] \\ &= \mathbb{E}_{t-1}[f(X_t)\Sigma_t^{-1}(X_t - \mu_t)] \\ &= \mathbb{E}_{t-1}[\nabla s_t(Z_t)] ,\end{aligned}$$

where in the first equality we used that Y_{t-1} and Σ_t are $\mathscr{F}_{t-1}$-measurable and $\mathbb{E}_{t-1}[X_t] = \mu_t$. In the second equality we substituted the definition of $Y_t = f(X_t) + \varepsilon_t$ and used the assumption that the noise is conditionally zero mean. The last follows from Lemma 3c. Part *(a)* follows by showing that $|\langle \mu_t, \mathcal{E}_1\rangle| \leq \lambda/n$, which we do in the next step. Moving now to the Hessian, the same reasoning yields

$$\begin{aligned}\mathbb{E}_{t-1}[H_t] = \lambda &\mathbb{E}_{t-1}[(Y_t - Y_{t-1})\Sigma_t^{-1/2}(W_t W_t^\top - \mathbb{1})\Sigma_t^{-1/2}] \\ &- \underbrace{\lambda \mathbb{E}_{t-1}[I_t (Y_t - Y_{t-1})\Sigma_t^{-1/2}(W_t W_t^\top - \mathbb{1})\Sigma_t^{-1/2}]}_{\mathcal{E}_2} .\end{aligned}$$

Repeating again the argument above with the first term,

$$\begin{aligned}\lambda \mathbb{E}_{t-1}[(Y_t - Y_{t-1})\Sigma_t^{-1/2}(W_t W_t^\top - \mathbb{1})\Sigma_t^{-1/2}] &= \lambda \mathbb{E}_{t-1}[f(X_t)\Sigma_t^{-1/2}(W_t W_t^\top - \mathbb{1})\Sigma_t^{-1/2}] \\ &= \mathbb{E}_{t-1}[\nabla^2 s_t(Z_t)] ,\end{aligned}$$

where we used Lemma 3d. Part *(b)* follows by bounding $|\operatorname{tr}(A\mathcal{E}_2)|$.

Tedious part Now we handle the error terms $\mathcal{E}_1$ and $\mathcal{E}_2$. On $\{2 \leq t \leq \tau\}$, by Definition 4a, $\Sigma_t \leq 2\Sigma_1 = \frac{2r^2}{d^2}\mathbb{1}$. Therefore

$$\frac{\mathrm{D_{max}}}{2} \geq \left(1 + 2\|\Sigma_t\|^{1/2}\right)\sqrt{\log(4\mathrm{P})}\,. \tag{6}$$

Using Definition Theorem 4c, $|\mathbb{E}_{t-1}[Y_t] - Y_{t-1}| \leq \frac{\mathrm{D_{max}}}{2}$. Therefore, by Lemma 15b, Fact 11c and Theorem 12a,

$$\begin{aligned}
\mathbb{P}_{t-1}(I_t) &\leq \mathbb{P}_{t-1}(|Y_t - Y_{t-1}| \geq \mathrm{D_{max}}) + \mathbb{P}_{t-1}(\|W_t\| \geq \mathrm{W_{max}}) \\
&\leq \mathbb{P}_{t-1}\left(|Y_t - \mathbb{E}_{t-1}[Y_t]| \geq \mathrm{D_{max}}/2\right) + \mathbb{P}_{t-1}(\|W_t\| \geq \mathrm{W_{max}}) \\
&\leq \mathbb{P}_{t-1}\left(|Y_t - \mathbb{E}_{t-1}[Y_t]| \geq \left(1 + 2\|\Sigma_t\|^{1/2}\right)\sqrt{\log(4\mathrm{P})}\right) + \mathbb{P}_{t-1}(\|W_t\| \geq \mathrm{W_{max}}) \\
&\leq \frac{1}{\mathrm{P}}\,.
\end{aligned}$$

We also need a crude bound on the moments of $Y_t - Y_{t-1}$. Again, using Definition 4c, Theorem 15b, and Equation (6) noting that $\mathrm{P} \geq 3$, we obtain

$$\begin{aligned}
\|Y_t - Y_{t-1}\|_{t-1,\psi_2} &\leq \|Y_t - \mathbb{E}_{t-1}[Y_t]\|_{t-1,\psi_2} + \frac{\mathrm{D_{max}}}{2\sqrt{\log(2)}} \\
&\leq 1 + 2\|\Sigma_t\|^{1/2} + \frac{\mathrm{D_{max}}}{2\sqrt{\log(2)}} \\
&\leq \frac{\mathrm{D_{max}}}{2\sqrt{\log(2\mathrm{P})}} + \frac{\mathrm{D_{max}}}{2\sqrt{\log(2)}} \\
&\leq \mathrm{D_{max}}\,.
\end{aligned}$$

Therefore, by Lemma 12c,

$$\mathbb{E}_{t-1}[(Y_t - Y_{t-1})^4]^{1/4} \leq 6\mathrm{D_{max}}\,.$$

Two applications of the Cauchy-Schwarz inequality yield

$$\begin{aligned}
|\langle \mu_t, \mathcal{E}_1\rangle| &= \left|\mathbb{E}_{t-1}[I_t\langle \mu_t, (Y_t - Y_{t-1})\Sigma_t^{-1}(X_t - \mu_t)\rangle]\right| \\
&\leq \mathbb{P}_{t-1}(I_t)^{1/4}\mathbb{E}_{t-1}[(Y_t - Y_{t-1})^4]^{1/4}\mathbb{E}_{t-1}[\langle \mu_t, \Sigma_t^{-1}(X_t - \mu_t)\rangle^2]^{1/2} \\
&\leq \frac{6\mathrm{D_{max}}\|\mu_t\|_t}{\mathrm{P}^{1/4}} \\
&\leq \frac{6\mathrm{D_{max}}\mathrm{F_{max}^{1/2}}}{\mathrm{P}^{1/4}} \\
&\leq \frac{\lambda}{n}\,. && \text{By choosing P large enough}
\end{aligned}$$

This completes the proof of part *(a)*. Repeating the Cauchy-Schwarz from the last step and letting $B = \Sigma_t^{-1/2} A \Sigma_t^{-1/2}$ and using Lemma 16c,

$$\begin{aligned}
|\operatorname{tr}(A\mathcal{E}_2)| &= \lambda \left| \mathbb{E}_{t-1}\left[I_t(Y_t - Y_{t-1}) \operatorname{tr}\left(\Sigma_t^{-1/2} A \Sigma_t^{-1/2}(W_t W_t^\top - \mathbb{1})\right)\right]\right| \\
&\leq \frac{6\lambda \mathrm{D}_{\max} \sqrt{(d^2 + 2d - 1)\operatorname{tr}(B^2)}}{\mathrm{P}^{1/4}} \\
&\leq \frac{6\lambda \mathrm{D}_{\max}(d+1)\operatorname{tr}(B)}{\mathrm{P}^{1/4}} \\
&= \frac{6\lambda \mathrm{D}_{\max}(d+1)\operatorname{tr}(\Sigma_t^{-1} A)}{\mathrm{P}^{1/4}} \\
&\leq \operatorname{tr}(\Sigma_t^{-1} A) \min\left(\frac{\lambda}{2n\mathrm{F}_{\max}}, \frac{\lambda}{dn}, \frac{1}{n\Sigma_{\max}^{-1}}\right),
\end{aligned}$$

where in the final inequality we chose P large enough.

Appendix C. Proof of Lemma 8

Since $\xi_z \in [\mathbf{0}, z] = \{\alpha z : \alpha \in [0,1]\}$, it follows from convexity that

$$\|\xi_z - \mu_t\|_{\Sigma_t^{-1}} \leq \max\left(\|z - \mu_t\|_{\Sigma_t^{-1}}, \|\mu_t\|_t\right).$$

By assumption $t \leq \tau$, which implies that $\frac{1}{2}\|\mu_t\|_t^2 \leq \mathrm{F}_{\max}$. Then, using the definition of λ,

$$\frac{\lambda}{1-\lambda}\|\mu_t\|_t \leq \frac{\lambda}{1-\lambda}\sqrt{2\mathrm{F}_{\max}} \leq \frac{\log(2)}{\sqrt{2\log(\mathrm{P})}}.$$

Recall that Z_t has law $\mathcal{N}(\mu_t, \beta^2 \Sigma_t)$ under $\mathbb{P}_{t-1}$. By Lemma 16de,

$$\mathbb{E}_{t-1}[\|Z_t\|^2] = \|\mu_t\|^2 + \beta^2 \operatorname{tr}(\Sigma_t) \qquad \text{and} \qquad \mathbb{E}_{t-1}[\|Z_t\|^4] \leq 3\left(\mathbb{E}_{t-1}[\|Z_t\|^2]\right)^2. \tag{7}$$

By Fact 11c and Lemma 12a, with probability at least $1 - 1/\mathrm{P}$,

$$\begin{aligned}
\frac{\lambda}{1-\lambda}\|Z_t - \mu_t\|_{\Sigma_t^{-1}} &\leq \frac{2\lambda\beta}{1-\lambda}\sqrt{d\log(2\mathrm{P})} \\
&= \frac{2}{1-\lambda}\sqrt{\lambda(2-\lambda)d\log(2\mathrm{P})} && \text{Since } \beta^2 = (2-\lambda)/\lambda \\
&\leq \frac{\log(2)}{\sqrt{2\log(2\mathrm{P})}},
\end{aligned}$$

where the second inequality follows by choosing c in the definition of λ small enough. This shows that

$$\mathbb{P}_{t-1}(Z_t \notin E) \leq 1/\mathrm{P} \tag{8}$$

for

$$E = \left\{ z \in \mathbb{R}^d : \frac{\lambda}{1-\lambda}\|z - \mu_t\|_{\Sigma_t^{-1}} \leq \frac{\log(2)}{\sqrt{2\log(\mathrm{P})}} \right\}.$$

By Lemma 3f, for $z \in E$ we have

$$\nabla^2 s_t(\xi_z) \geq \frac{1}{2}\nabla^2 s_t(\mu_t) - \frac{\|\Sigma_t^{-1/2}\|}{2\mathrm{P}}\mathbb{1}\,.$$

Therefore,

$$\begin{aligned}\mathbb{E}_{t-1}\left[\|Z_t\|^2_{\nabla^2 s_t(\xi_{Z_t})}\right] &\geq \frac{1}{2}\mathbb{E}_{t-1}\left[\mathbf{1}(Z_t \in E)\|Z_t\|^2_{\nabla^2 s_t(\mu_t)}\right] - \frac{\|\Sigma_t^{-1/2}\|}{2\mathrm{P}}\mathbb{E}[\|Z_t\|^2] \\ &= \frac{1}{2}\mathbb{E}_{t-1}\left[\|Z_t\|^2_{\nabla^2 s_t(\mu_t)}\right] - \frac{1}{2}\mathbb{E}_{t-1}\left[\mathbf{1}(Z_t \notin E)\|Z_t\|^2_{\nabla^2 s_t(\mu_t)}\right] - \frac{\|\Sigma_t^{-1/2}\|}{2\mathrm{P}}\mathbb{E}[\|Z_t\|^2]\,. \end{aligned} \tag{9}$$

The last term in Equation (9) is bounded using Equation (7):

$$\begin{aligned}\frac{\mathbb{E}_{t-1}[\|Z_t\|^2]\|\Sigma_t^{-1/2}\|}{2\mathrm{P}} &\leq \frac{\mathbb{E}_{t-1}[\|Z_t\|^2]\sqrt{\Sigma_{\max}^{-1}}}{2\mathrm{P}} \\ &= \frac{\left(\|\mu_t\|^2 + \beta^2\operatorname{tr}(\Sigma_t)\right)\sqrt{\Sigma_{\max}^{-1}}}{2\mathrm{P}} \\ &\leq \frac{\left(2\|\Sigma_1\|\mathrm{F}_{\max} + 2\beta^2\operatorname{tr}(\Sigma_1)\right)\sqrt{\Sigma_{\max}^{-1}}}{2\mathrm{P}} \\ &\leq \frac{1}{2n}\,, \end{aligned} \tag{10}$$

where the last inequality follows by choosing P large enough and in the first and second inequalities we used the facts that on $\{2 \leq t \leq \tau\}$,

$$\|\Sigma_t^{-1/2}\| \leq \sqrt{\operatorname{tr}(\Sigma_t^{-1})} \leq \sqrt{\Sigma_{\max}^{-1}} \qquad \text{and} \qquad \|\mu_t\|^2 \leq \|\Sigma_t\|\|\mu_t\|_t^2 \leq 2\|\Sigma_1\|\mathrm{F}_{\max}\,.$$

For the second to last term in Equation (9), Theorem 3e, Equation (7) and Equation (8) imply

$$\begin{aligned}\frac{1}{2}\mathbb{E}_{t-1}\left[\mathbf{1}(Z_t \notin E)\|Z_t\|^2_{\nabla^2 s_t(\mu_t)}\right] &\leq \frac{1}{2}\|\Sigma_t^{-1/2}\|\mathbb{E}_{t-1}\left[\mathbf{1}(Z_t \notin E)\|Z_t\|^2\right] \\ &\leq \frac{1}{2}\|\Sigma_t^{-1/2}\|\sqrt{\mathbb{P}_{t-1}(Z_t \notin E)\mathbb{E}_{t-1}\left[\|Z_t\|^4\right]} \\ &\leq \mathbb{E}_{t-1}\left[\|Z_t\|^2\right]\|\Sigma_t^{-1/2}\|\sqrt{\frac{3}{4\mathrm{P}}} \\ &\leq \frac{1}{2n}\,, \end{aligned}$$

where the last inequality follows the same way as Equation (10), again making sure that P is chosen large enough. Therefore,

$$\mathbb{E}_{t-1}\left[\|Z_t\|^2_{\nabla^2 s_t(\xi_{Z_t})}\right] \geq \frac{1}{2}\mathbb{E}_{t-1}\left[\|Z_t\|^2_{\nabla^2 s_t(\mu_t)}\right] - \frac{1}{n}\,.$$

Part *(b)* follows along the same lines. Using Theorem 3ef,

$$\begin{aligned}
\frac{1}{2}\mathbb{E}_{t-1}\left[\|\mu_t\|^2_{\nabla^2 s_t(Z_t)}\right] &= \frac{1}{2}\mathbb{E}_{t-1}\left[\mathbf{1}(Z_t \in E)\|\mu_t\|^2_{\nabla^2 s_t(Z_t)}\right] + \frac{1}{2}\mathbb{E}_{t-1}\left[\mathbf{1}(Z_t \notin E)\|\mu_t\|^2_{\nabla^2 s_t(Z_t)}\right] \\
&\leq \mathbb{E}_{t-1}\left[\|\mu_t\|^2_{\nabla^2 s_t(\mu_t)}\right] + \frac{\|\Sigma_t^{-1/2}\|}{2\mathrm{P}}\mathbb{E}_{t-1}\left[\|\mu_t\|^2\right] \\
&\quad + \frac{\|\Sigma_t^{-1/2}\|}{2}\mathbb{E}_{t-1}\left[\mathbf{1}(Z_t \notin E)\|\mu_t\|^2\right] \\
&\leq \mathbb{E}_{t-1}\left[\|\mu_t\|^2_{\nabla^2 s_t(\mu_t)}\right] + \frac{\|\Sigma_t^{-1/2}\|\|\Sigma_t\|\mathrm{F}_{\max}}{\mathrm{P}} \\
&\leq \mathbb{E}_{t-1}\left[\|\mu_t\|^2_{\nabla^2 s_t(\mu_t)}\right] + \frac{1}{n}\,,
\end{aligned}$$

where in the the second to last inequality we used the independence of Z_t and μ_t under $\mathbb{P}_{t-1}$ and Equation (8), while the final inequality we again used that on $\{2 \leq t \leq \tau\}$ both $\|\Sigma_t^{-1/2}\|$ and $\|\Sigma_t\|$ are bounded by polynomials in d, n and r.

Appendix D. Proof of Theorem 5

By definition $E_t = E_t^{(a)} \cap E_t^{(b)} \cap E_t^{(c)}$, where

$$\begin{aligned}
E_t^{(a)} &= \{\Sigma_{t+1} \leq 2\Sigma_1, \Sigma_{t+1} \leq 2\Sigma_t\}\,, \\
E_t^{(b)} &= \{\mathrm{tr}(\Sigma_{t+1}^{-1}) \leq \Sigma_{\max}^{-1}\}\,, \\
E_t^{(c)} &= \{|\mathbb{E}_t[Y_{t+1}] - Y_t| \leq \mathrm{D}_{\max}/2\}\,.
\end{aligned}$$

The plan is to show that all of these events occur with high probability; then a naive application of the union bound finishes the proof.

Step 1: Stability We start by showing that the mean and covariance change slowly, which is a consequence of the truncation in the algorithm and the choice of parameters. For every $t \geq 2$,

$$\Sigma_{t+1}^{-1} = \Sigma_t^{-1} + \frac{\eta}{4}H_t = \Sigma_t^{-1/2}\left(\mathbb{1} + \frac{\eta}{4}\Sigma_t^{1/2}H_t\Sigma_t^{1/2}\right)\Sigma_t^{-1/2}\,. \tag{11}$$

By the definition of H_t,

$$\begin{aligned}
\left\|\frac{\eta}{4}\Sigma_t^{1/2}H_t\Sigma_t^{1/2}\right\| &= \left\|\frac{\eta\lambda D_t}{4}(W_tW_t^\top - \mathbb{1})\right\| && (12) \\
&\leq \frac{\eta\lambda \mathrm{D}_{\max}(1+\mathrm{W}_{\max}^2)}{4} && (13) \\
&\leq \frac{1}{2d}\,, && \text{Definitions of } \eta \text{ and } \lambda
\end{aligned}$$

where the last inequality follows by choosing c, C and P sufficiently small, large and large, respectively. Note that this implies that $\Sigma_{t+1} \leq 2\Sigma_t$ for all $1 \leq t \leq n$. Moving now to the mean, on the event $\{2 \leq t \leq \tau\}$ we have

$$\begin{aligned}
\|\mu_{t+1} - \mu_t\| &= \eta\|\Sigma_{t+1} g_t\| \\
&= \eta\|D_t \Sigma_{t+1} \Sigma_t^{-1}(X_t - \mu_t)\| \\
&\leq \eta \mathrm{D}_{\max} \mathrm{W}_{\max} \|\Sigma_{t+1}\Sigma_t^{-1/2}\| \\
&\leq \sqrt{2}\eta \mathrm{D}_{\max} \mathrm{W}_{\max} \|\Sigma_{t+1}^{1/2}\| \\
&\leq 2\eta \mathrm{D}_{\max} \mathrm{W}_{\max} \|\Sigma_1^{1/2}\| \\
&\leq \frac{r}{d}\,. \qquad (14)
\end{aligned}$$

Step 2: Baseline quality The plan in this step is to show that on $\{2 \leq t \leq \tau\}$,

$$\mathbb{P}_{t-1}\left(|\mathbb{E}_t[Y_{t+1}] - Y_t| \leq \frac{\mathrm{D}_{\max}}{2}\right) \leq \frac{1}{n^2}\,. \qquad (15)$$

and then use a union bound to establish that $\mathbb{P}(\cap_{t=1}^{\tau} E_t^{(c)}) \geq 1 - 1/n$. There are two parts to establishing Equation (15):

(1) Showing that Y_t is close to $\mathbb{E}_{t-1}[Y_t]$, which follows from concentration for Lipschitz functions and the definition of the noise model.

(2) Showing that $\mathbb{E}_{t-1}[Y_t]$ is close to $\mathbb{E}_t[Y_{t+1}]$, which is a consequence of the stability of the algorithm that was shown in the previous step.

We start with the second. By Lemma 10ab and using that $\frac{1}{4}\eta\|\Sigma_t^{1/2} H_t \Sigma_t^{1/2}\| \leq \frac{1}{2d}$,

$$\frac{2d-1}{2d}\Sigma_t^{-1} \leq \Sigma_{t+1}^{-1} \leq \frac{2d+1}{2d}\Sigma_t^{-1} \quad \text{and so} \quad \frac{2d}{2d+1}\Sigma_t \leq \Sigma_{t+1} \leq \frac{2d}{2d-1}\Sigma_t\,.$$

Because f is Lipschitz, if the mean and covariance matrices change slowly from one round to the next, then the mean loss should also change slowly. This phenomenon is captured by Lemma 14, which yields

$$\begin{aligned}
|\mathbb{E}_t[Y_{t+1}] - \mathbb{E}_{t-1}[Y_t]| &\leq \sqrt{\|\mu_{t+1} - \mu_t\|^2 + \operatorname{tr}\left(\Sigma_t + \Sigma_{t+1} - 2\left(\Sigma_t^{1/2}\Sigma_{t+1}\Sigma_t^{1/2}\right)^{1/2}\right)} \\
&\leq \sqrt{\frac{r^2}{d^2} + \operatorname{tr}\left((1 - \sqrt{1 - 1/(2d)})\Sigma_t + (1 - \sqrt{1 - 1/(2d)})\Sigma_{t+1}\right)} \\
&\leq \sqrt{\frac{r^2}{d^2} + \frac{1}{2d}\operatorname{tr}\left(\Sigma_t + \Sigma_{t+1}\right)} \\
&\leq \frac{3r}{d}\,,
\end{aligned}$$

where the second inequality follows by Equation (14)-Equation (12) and standard monotonicity properties of the trace.

On $\{2 \leq t \leq \tau\}$, $E_{t-1}^{(a)}$ holds and so $\|\Sigma_t\| \leq 2\|\Sigma_1\| = \frac{2r^2}{d^2}$ and by Theorem 15b and Theorem 12a,

$$\mathbb{P}_{t-1}\left(|Y_t - \mathbb{E}_{t-1}[Y_t]| \geq \left(1 + \frac{2r}{d}\right)\sqrt{2\log(2n^2)}\right) \leq \frac{1}{n^2}\,.$$

Hence, combining the above two displayed inequalities, it follows that with $\mathbb{P}_{t-1}$-probability at least $1 - 1/n^2$,

$$|\mathbb{E}_t[Y_{t+1}] - Y_t| \leq |\mathbb{E}_t[Y_{t+1}] - \mathbb{E}_{t-1}[Y_t]| + |\mathbb{E}_{t-1}[Y_t] - Y_t| \leq \left(1 + \frac{5r}{d}\right)\sqrt{2\log(2n^2)} \leq \frac{\mathrm{D}_{\max}}{2}\,,$$

where the last inequality follows by choosing the constants in the definition of $\mathrm{D}_{\max}$ large enough. By a union bound it now follows that $\mathbb{P}(\cap_{t=1}^{\tau} E_t^{(c)}) \geq 1 - 1/n$.

Step 3: Lower bound on covariance In this step we show that with probability at least $1 - 1/n$

$$\mathrm{tr}\left(\Sigma_{\tau+1}^{-1}\right) < \Sigma_{\max}^{-1}\,,$$

which implies that $\mathbb{P}(\cap_{t=1}^{\tau} E_t^{(\mathrm{b})}) = \mathbb{P}\left(\mathrm{tr}(\Sigma_{\tau+1}^{-1}) \leq \Sigma_{\max}^{-1}\right) \geq 1 - 1/n$. Since Σ_t^{-1} is positive definite for all t, by Markov's inequality,

$$\begin{aligned}
\mathbb{P}\left(\mathrm{tr}(\Sigma_{\tau+1}^{-1}) \geq \Sigma_{\max}^{-1}\right) &\leq \frac{1}{\Sigma_{\max}^{-1}}\mathbb{E}\left[\mathrm{tr}(\Sigma_{\tau+1}^{-1})\right] \\
&= \frac{1}{\Sigma_{\max}^{-1}}\left(\mathrm{tr}(\Sigma_1^{-1}) + \frac{\eta}{4}\mathbb{E}\left[\sum_{t=2}^{n}\mathrm{tr}(H_t)\mathbf{1}_{\tau\geq t}\right]\right) \\
&\leq \frac{1}{\Sigma_{\max}^{-1}}\left(\frac{d^3}{r^2} + \frac{\eta}{4}\mathbb{E}\left[\sum_{t=2}^{n}\mathrm{tr}(\nabla^2 s_t(Z_t))\mathbf{1}_{\tau\geq t}\right] + \frac{\eta}{4}\right) && \text{Theorem 6b} \\
&\leq \frac{1}{\Sigma_{\max}^{-1}}\left(\frac{d^3}{r^2} + \frac{\eta n d\sqrt{\Sigma_{\max}^{-1}}}{4} + \frac{\eta}{4}\right) && \text{Theorem 3e} \\
&\leq \frac{1}{\sqrt{\Sigma_{\max}^{-1}}}\left(\frac{d^2}{r^2} + \frac{nd}{4} + 1\right) \\
&\leq \frac{1}{n}\,,
\end{aligned}$$

where the second to last inequality follows by naive simplification and the last using the definition of $\Sigma_{\max}^{-1}$ in Theorem 4b.

Step 4: Upper bound on covariance It remains to show that $\Sigma_{\tau+1} \leq 2\Sigma_1$ with high probability, which we do by showing that Σ_{t+1}^{-1} is unlikely to decrease too much. The sphere embedded in $\mathbb{R}^d$ is denoted by $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$ and let $x \in \mathbb{S}^{d-1}$ be an arbitrary unit vector. The update of H_t guarantees that, for any $s \leq \tau$,

$$\|x\|_{s+1}^2 = \|x\|_2^2 + \sum_{t=2}^{s}\frac{1}{4}\eta\|x\|_{H_t}^2\,,$$

where you should note that $\|x\|_2^2 = \|x\|_{\Sigma_1^{-1}}^2$ is not the 2-norm. Therefore,

$$\log\left(\frac{\|x\|_{\tau+1}^2}{\|x\|_2^2}\right) = \sum_{t=2}^{\tau}\log\left(1 + \frac{\frac{1}{4}\eta\|x\|_{H_t}^2}{\|x\|_t^2}\right)\,.$$

Next we lower bound the terms in the sum. Fix some $t \leq \tau$. By Fact 11b and simplifying,

$$\left\| \frac{\frac{1}{4}\eta\|x\|_{H_t}^2}{\|x\|_t^2} \right\|_{t-1,\psi_1} = \frac{\eta\lambda}{4\|x\|_t^2} \left\| D_t \|\Sigma_t^{-1/2} x\|_{W_t W_t^\top - \mathbb{1}}^2 \right\|_{t-1,\psi_1} \leq \frac{5}{4}\eta\lambda \mathrm{D}_{\max} .$$

Since $\mathbb{E}_{t-1}[H_t]$ is very close to the Hessian of a convex function, we should expect that $\mathbb{E}_{t-1}[\|x\|_{H_t}^2]$ is nearly positive. Indeed, by Lemma 6b,

$$\mathbb{E}_{t-1}\left[\frac{\frac{1}{4}\eta\|x\|_{H_t}^2}{\|x\|_t^2}\right] \geq \frac{\mathbb{E}_{t-1}\left[\|x\|_{\nabla^2 s(Z_t)}^2\right]}{\|x\|_t^2} - \frac{\lambda}{dn} \geq -\frac{1}{10n} .$$

Let $\mathcal{C} \subset \mathbb{S}^{d-1}$ be a finite cover of the sphere such that for all $y \in \mathbb{S}^{d-1}$ there exists an $x \in \mathcal{C}$ with $\|x - y\| \leq 1/\mathrm{P}$. By Corollary 4.2.13 of Vershynin (2018), the cover can be chosen so that

$$|\mathcal{C}| \leq (2\mathrm{P} + 1)^d .$$

Next, by Bernstein's inequality (Lemma 13), with probability at least $1 - 1/n$, for all $x \in \mathcal{C}$,

$$\sum_{t=2}^{\tau} \frac{\frac{1}{4}\eta\|x\|_{H_t}^2}{\|x\|_t^2} \geq -83\eta\lambda\mathrm{D}_{\max}\sqrt{n\log(n|\mathcal{C}|)} - \frac{1}{10} .$$

Furthermore, by Lemma 12b, with probability at least $1 - 1/n$, for all $x \in \mathcal{C}$ and $t \leq \tau$,

$$\left| \frac{\frac{1}{4}\eta\|x\|_{H_t}^2}{\|x\|_t^2} \right| \leq \eta\lambda\mathrm{D}_{\max}\log(2n^2|\mathcal{C}|) \leq 1 .$$

Therefore, using $\log(1 + x) \geq x - x^2$ for $x \geq -1$, we obtain

$$\begin{aligned}
\log\left(\frac{\|x\|_{\tau+1}^2}{\|x\|_2^2}\right) &= \sum_{t=2}^{\tau} \log\left(1 + \frac{\frac{1}{4}\eta\|x\|_{H_t}^2}{\|x\|_t^2}\right) \\
&\geq \sum_{t=2}^{\tau}\left[\frac{\frac{1}{4}\eta\|x\|_{H_t}^2}{\|x\|_t^2} - \left(\frac{\frac{1}{4}\eta\|x\|_{H_t}^2}{\|x\|_t^2}\right)^2\right] \\
&\geq -83\eta\lambda\mathrm{D}_{\max}\sqrt{n\log(n|\mathcal{C}|)} - n\eta^2\lambda^2\mathrm{D}_{\max}^2\log^2(2n^2|\mathcal{C}|) - \frac{1}{10} \\
&\geq -\frac{1}{20} > -\log(4/3) .
\end{aligned}$$

Combining the above calculations with the analysis in the previous step and a union bound shows that with probability at least $1 - 3/n$ it holds that $\|\Sigma_{\tau+1}^{-1}\| \leq \Sigma_{\max}^{-1}$ and for all $x \in \mathcal{C}$,

$$\|x\|_{\tau+1}^2 \geq \frac{3\|x\|_2^2}{4} = \frac{3}{4}\|x\|_{\Sigma_1^{-1}}^2 = \frac{3d^2}{4r^2} .$$

On this event,

$$\begin{aligned}
\min_{y \in \mathbb{S}^{d-1}} \|y\|_{\tau+1} &= \min_{y \in \mathbb{S}^{d-1}} \max_{x \in \mathcal{C}} \|y - x + x\|_{\tau+1} \\
&\geq \min_{y \in \mathbb{S}^{d-1}} \max_{x \in \mathcal{C}} \left(\|x\|_{\tau+1} - \|x - y\|_{\tau+1}\right) \\
&\geq \min_{y \in \mathbb{S}^{d-1}} \left(\min_{x \in \mathcal{C}} \|x\|_{\tau+1} - \min_{x \in \mathcal{C}} \|\Sigma_{\tau+1}^{-1}\| \|x - y\|\right) \\
&\geq \min_{x \in \mathcal{C}} \|x\|_{\tau+1} - \frac{\|\Sigma_{\tau+1}^{-1}\|}{\mathrm{P}} \\
&\geq \sqrt{\frac{3d^2}{4r^2}} - \frac{\Sigma_{\max}^{-1}}{\mathrm{P}} \\
&\geq \sqrt{\frac{d^2}{2r^2}}\,,
\end{aligned}$$

where the last inequality holds for P large enough. Therefore with probability at least $1 - 3/n$, $\Sigma_{\tau+1}^{-1} \geq \Sigma_1^{-1}/2$, which implies that $\Sigma_{\tau+1} \leq 2\Sigma_1$ and so, combined with the fact that $\Sigma_{t+1} \leq 2\Sigma_t$ holds for all t,

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{t=1}^{\tau} \left(E_t^{(a)} \cap E_t^{(b)}\right)\right) &= \mathbb{P}\left(\Sigma_{\tau+1} \leq 2\Sigma_1 \text{ and } \operatorname{tr}(\Sigma_{\tau+1}^{-1}) \leq \Sigma_{\max}^{-1} \text{ and } \Sigma_{t+1} \leq 2\Sigma_t \text{ for } 1 \leq t \leq n\right) \\
&\geq 1 - 3/n\,.
\end{aligned}$$

Appendix E. Technical lemmas

Lemma 10 *Suppose that A, B and C are square matrices with A and B positive definite, C symmetric and $\|C\| \leq \varepsilon \leq 1$. Then*

(a) $A^{1/2}(\mathbb{1} + C)A^{1/2} \geq (1 - \varepsilon)A$ and $A^{1/2}(\mathbb{1} + C)A^{1/2} \leq (1 + \varepsilon)A$.

(b) If $A \leq B$, then $B^{-1} \leq A^{-1}$.

Appendix F. Concentration bounds

None of the results in this section are novel in any way. In some cases we needed to include explicit constants where published results simplify with unspecified universal constants. We are expedient in our calculation of these constants. In case you wanted a truly refined analysis, then the Orlicz-norm style analysis should be replaced with the kind of analysis that relies on moment-generating functions.

Fact 11 *Let $W \stackrel{d}{=} \mathcal{N}(\mathbf{0}, \mathbb{1})$. Then*

(a) $\|\langle x, W\rangle\|_{\psi_2} = 2\sqrt{2/3}\|x\| \leq 2\|x\|$ for any $x \in \mathbb{R}^d$;

(b) $\|\operatorname{tr}(AWW^\top)\|_{\psi_1} \leq 3\operatorname{tr}(A)$ for any positive semidefinite $A \in \mathbb{R}^{d \times d}$;

(c) $\big\|\|W\|\big\|_{\psi_2}^2 = \big\|\|W\|^2\big\|_{\psi_1} \leq 8d/3$;

(d) $\big\|\|WW^\top - \mathbb{1}\|\big\|_{\psi_1} \leq 5d$.

Proof All results follow from explicit calculation using the Gaussian density:

(a) Since $\|\langle x, W\rangle\|_{\psi_2} = \|x\| \big\| \langle x, W\rangle / \|x\| \big\|_{\psi_2}$, we may assume without loss of generality that $\|x\| = 1$ so that $\langle x, W\rangle \overset{d}{=} \mathcal{N}(0, 1)$. Then

$$\mathbb{E}[\exp(\langle x, W\rangle^2/t^2)] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(-z^2/2 + z^2/t^2)\, \mathrm{d}z = 1/\sqrt{1 - 2/t^2}\,.$$

The right-hand side is less than 2 for $t \geq 2\sqrt{2/3} \approx 1.63$.

(b) Let $A = U^{-1}\Lambda U$ for orthonormal U and Λ a diagonal matrix with eigenvalues $\lambda_1, \ldots, \lambda_d$. By rotational invariance UW and W have the same distribution and therefore so do $\operatorname{tr}(AWW^\top)$ and $\|W\|_\Lambda^2$. Therefore $\|\operatorname{tr}(AWW^\top)\|_{\psi_1} = \|\sum_{m=1}^d \lambda_m W_m^2\|_{\psi_1} \leq \sum_{m=1}^d \lambda_m \|W_m^2\|_{\psi_1}$. The result follows since, by part *(a)*, $\|W_m^2\|_{\psi_1} = \|W_m\|_{\psi_2}^2 = 8/3 \leq 3$.

(c) Using *(b)*, $\big\|\|W\|\big\|_{\psi_2}^2 = \big\|\|W\|^2\big\|_{\psi_1} = \|\operatorname{tr}(\mathbb{1}WW^\top)\|_{\psi_1} \leq 8d/3$.

(d) Using *(b)*, $\big\|\|WW^\top - \mathbb{1}\|\big\|_{\psi_1} \leq \big\|1 + \|WW^\top\|\big\|_{\psi_1} = \log(2) + \|\operatorname{tr}(WW^\top)\|_{\psi_1} \leq 5d$.

■

Lemma 12 (Proposition 2.5.2, Proposition 2.7.1, Vershynin 2018) *Let X be a real random variable. Then for all $t \geq 0$,*

(a) $\mathbb{P}(|X| \geq t) \leq 2\exp\left(-t^2/\|X\|_{\psi_2}^2\right)$*;*

(b) $\mathbb{P}(|X| \geq t) \leq 2\exp\left(-t/\|X\|_{\psi_1}\right)$*;*

(c) $\mathbb{E}[|X|^p]^{1/p} \leq 3\sqrt{p}\|X\|_{\psi_2}$ *for all $p \geq 1$;*

Lemma 13 *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and $X_1, \ldots, X_n$ be a sequence of random variables adapted to a filtration $(\mathscr{F}_t)_{t=1}^n$ and let $\|X_t\|_{t-1,\psi_1}$ be the $\|\cdot\|_{\psi_1}$ norm of X_t with respect to $\mathbb{P}(\cdot|\mathscr{F}_{t-1})$. Suppose that τ is a stopping time with respect to $(\mathscr{F}_t)_{t=1}^n$ and $\|X_t\|_{t-1,\psi_1} \leq \alpha$ almost surely on $\{t - 1 < \tau\}$. Then for any $\delta \in (0, 1)$,*

$$\mathbb{P}\left(\sum_{t=1}^{\tau} X_t - \mathbb{E}[X_t|\mathscr{F}_{t-1}] \geq 66 \max\left[\alpha\sqrt{n\log(1/\delta)}\,,\, \alpha\log(1/\delta)\right]\right) \leq \delta\,.$$

Proof Repeat the proof of the standard Bernstein inequality to the sequence $(Y_t)_{t=1}^n$ with $Y_t = X_t\mathbf{1}(t \leq \tau)$ (Vershynin, 2018, Theorem 2.8.1). ■

Lemma 14 *Let $X \overset{d}{=} \mathcal{N}(\mu_1, \Sigma_1)$ and $Y \overset{d}{=} \mathcal{N}(\mu_2, \Sigma_2)$ and $f : \mathbb{R}^d \to \mathbb{R}$ be Lipschitz. Then*

$$|\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]| \leq \sqrt{\|\mu_1 - \mu_2\|^2 + \operatorname{tr}\left(\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right)}\,.$$

Proof Let $W_k(p, q)$ be the k-Wasserstein distance between probability measures p and q for $k \in \{1, 2\}$. Since f is Lipschitz, $|\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]| \leq W_1(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2))$ and by convexity $W_1 \leq \sqrt{W_2}$. Therefore using the closed form of W_2 between Gaussians by Dowson and Landau (1982) yields

$$
\begin{aligned}
|\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]| &\leq \sqrt{W_2(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2))} \\
&= \sqrt{\|\mu_1 - \mu_2\|^2 + \operatorname{tr}\left(\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right)}\,.
\end{aligned}
$$

■

Lemma 15 *Suppose that X has law $\mathcal{N}(\mu, \Sigma)$ and $\|\varepsilon\|_{\mathbb{P}(\cdot|X),\psi_2} \leq 1$. Then with $Y = f(X) + \varepsilon$ and any $\delta \in (0, 1)$,*

(a) $\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq 2\|\Sigma\|^{1/2}$*; and*

(b) $\|Y - \mathbb{E}[Y]\|_{\psi_2} \leq 1 + 2\|\Sigma\|^{1/2}$.

Proof Let W have law $\mathcal{N}(\mathbf{0}, \mathbb{1})$ and $h(w) = f(\Sigma^{1/2}w + \mu)$. Since f is Lipschitz, $h(u) - h(v) \leq \|\Sigma^{1/2}(u - v)\| \leq \|\Sigma\|^{1/2}\|u - v\|$. Since this holds for all u and v, h is $\|\Sigma\|^{1/2}$-Lipschitz with respect to the Euclidean norm. By Theorem 5.6 of Boucheron et al. (2013),

$$
\mathbb{P}\left(|f(X) - \mathbb{E}[f(X)]| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\|\Sigma\|}\right)\,.
$$

Therefore, by Proposition 2.5.2 of Vershynin (2018), $\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq 2\|\Sigma\|^{1/2}$, which establishes *(a)*. By the triangle inequality,

$$
\|Y - \mathbb{E}[Y]\|_{\psi_2} = \|f(X) + \varepsilon - \mathbb{E}[f(X)]\|_{\psi_2} \leq 2\|\Sigma\|^{1/2} + 1\,,
$$

which yields *(b)*. ■

Lemma 16 *Let W have law $\mathcal{N}(\mathbf{0}, \mathbb{1})$ and A be positive definite and Z have law $\mathcal{N}(\mu, \Sigma)$. Then the following hold:*

(a) $\mathbb{E}[\|W\|_A^2] = \operatorname{tr}(A)$ *and* $\mathbb{E}[\|W\|_A^4] = \operatorname{tr}(A)^2 + 2\operatorname{tr}(A^2)$.

(b) $\mathbb{E}[\operatorname{tr}(A(WW^\top - \mathbb{1})A(WW^\top - \mathbb{1}))] = \operatorname{tr}(A)^2 + \operatorname{tr}(A^2)$.

(c) $\mathbb{E}[\operatorname{tr}(A(WW^\top - \mathbb{1})(WW^\top - \mathbb{1})A)] = (d^2 + 2d - 1)\operatorname{tr}(A^2)$.

(d) $\mathbb{E}[\|Z\|^2] = \operatorname{tr}(\Sigma) + \|\mu\|^2$.

(e) $\mathbb{E}[\|Z\|^4] = \operatorname{tr}(\Sigma)^2 + 2\operatorname{tr}(\Sigma^2) + \|\mu\|_\Sigma^2 + \|\mu\|^4 + 2\operatorname{tr}(\Sigma)\|\mu\|^2 \leq 3(\mathbb{E}[\|Z\|^2])^2$.

Proof By rotational invariance of the Gaussian and because A is diagonalised by a rotation matrix it suffices to consider the case that A is diagonal with eigenvalues $(\lambda_m)_{m=1}^d$. Then

$$
\mathbb{E}[\|W\|_A^2] = \mathbb{E}\left[\sum_{m=1}^d \lambda_m W_m^2\right] = \sum_{m=1}^d \lambda_m = \operatorname{tr}(A)\,.
$$

Furthermore, denoting the eigenvalues of A by $\lambda_1, \dots, \lambda_d$, we have

$$\mathbb{E}[\|W\|_A^4] = \mathbb{E}\left[\left(\sum_{m=1}^d \lambda_m W_m^2\right)^2\right] = \operatorname{tr}(A)^2 + \sum_{m=1}^d \lambda_m^2(\mathbb{E}[W_m^4] - 1) = \operatorname{tr}(A)^2 + 2\operatorname{tr}(A^2)\,,$$

where in the final equality we used the fact that the fourth moment of a standard Gaussian is $\mathbb{E}[W_m^4] = 3$. For *(b)*,

$$\begin{aligned}
\mathbb{E}[\operatorname{tr}(A(WW^\top - \mathbb{1})A(WW^\top - \mathbb{1}))] &= \mathbb{E}[\|W\|_A^4 - 2\operatorname{tr}(AWW^\top A) + \operatorname{tr}(A^2)] \\
&= \mathbb{E}[\|W\|_A^4] - \operatorname{tr}(A^2) \\
&= \operatorname{tr}(A)^2 + \operatorname{tr}(A^2)\,,
\end{aligned}$$

where in the final equality we used part *(a)*. Part *(c)* follows similarly:

$$\begin{aligned}
\mathbb{E}[\operatorname{tr}(A(WW^\top - \mathbb{1})(WW^\top - \mathbb{1})A)] &= \mathbb{E}[\|W\|^4 \operatorname{tr}(A^2) - 2\operatorname{tr}(AWW^\top A) + \operatorname{tr}(A^2)] \\
&= (d^2 + 2d - 1)\operatorname{tr}(A^2)\,,
\end{aligned}$$

where we used the second statement of part *(a)* in the last step. Parts *(d)* and *(e)* follow by noting that Z and $\Sigma^{1/2}W + \mu$ have the same law and using parts *(a)* and *(b)*. In particular,

$$\mathbb{E}[\|Z\|^2] = \mathbb{E}[\|\Sigma^{1/2}W + \mu\|^2] = \mathbb{E}[\|W\|_\Sigma^2] + \|\mu\|^2 = \operatorname{tr}(\Sigma) + \|\mu\|^2\,.$$

Furthermore,

$$\begin{aligned}
\mathbb{E}[\|Z\|^4] &= \mathbb{E}\left[\left((W^\top\Sigma^{1/2} + \mu^\top)(\Sigma^{1/2}W + \mu)\right)^2\right] \\
&= \mathbb{E}\left[\left(\|W\|_\Sigma^2 + 2W^\top\Sigma^{1/2}\mu + \|\mu\|^2\right)^2\right] \\
&= \mathbb{E}\left[\|W\|_\Sigma^4 + \mu^\top\Sigma^{1/2}WW^\top\Sigma^{1/2}\mu + \|\mu\|^4 + 2\|W\|_\Sigma^2\|\mu\|^2\right] \\
&= \operatorname{tr}(\Sigma)^2 + 2\operatorname{tr}(\Sigma^2) + \|\mu\|_\Sigma^2 + \|\mu\|^4 + 2\operatorname{tr}(\Sigma)\|\mu\|^2 \\
&\leq \operatorname{tr}(\Sigma)^2 + 2\operatorname{tr}(\Sigma^2) + \|\mu\|^4 + 3\operatorname{tr}(\Sigma)\|\mu\|^2 \\
&\leq 1.5(\operatorname{tr}(\Sigma) + \|\mu\|^2)^2 + 1.5\operatorname{tr}(\Sigma)^2 \\
&\leq 3(\mathbb{E}[\|Z\|^2])^2,
\end{aligned}$$

where in the last step we used part *(d)*. ■